STEPHEN W. HAWKING

THE ILLUSTRATED THEORY OF EVERYTHING

The Theory of Everything: The Origin and Fate of the Universe

The Theory of Everything: The Origin and Fate of the Universe

The Theory of Everything: The Origin and Fate of the Universe

By Hawking, Stephen

INTRODUCTION

In this series of lectures I shall try to give an outline of what we think is thehistory of the universe from the big bang to black holes. In the first lectureI shall briefly review past ideas about the universe and how we got to ourpresent picture. One might call this the history of the history of the universe.

In the second lecture I shall describe how both Newton's and Einstein's the-ories of gravity led to the conclusion that the universe could not be static; ithad to be either expanding or contracting. This, in turn, implied that theremust have been a time between ten and twenty billion years ago when thedensity of the universe was infinite. This is called the big bang. It would havebeen the beginning of the universe.

In the third lecture I shall talk about black holes. These are formed when amassive star or an even larger body collapses in on itself under its owngravitational pull. According to Einstein's general theory of relativity, anyonefoolish enough to fall into a black hole will be lost forever. They will not beable to come out of the black hole again. Instead, history, as far as they areconcerned, will come to a sticky end at a singularity. However, generalrelativity is a classical theory-that is, it doesn't take into account theuncertainty principle of quantum mechanics.

In the fourth lecture I shall describe how quantum mechanics allows energy toleak out of black holes. Black holes aren't as black as they are painted.In the fifth lecture I shall apply quantum mechanical ideas to the big bang andthe origin of the universe. This leads to the idea that space-time may be finitein extent but without boundary or edge. It would be like the surface of theEarth but with two more dimensions.In the sixth lecture I shall show how this new boundary proposal could explainwhy the past is so different

from the future, even though the laws of physics aretime symmetric.

Finally, in the seventh lecture I shall describe how we are trying to find aunified theory that will include quantum mechanics, gravity, and all the otherinteractions of physics. If we achieve this, we shall really understand theuniverse and our position in it.

The Theory of Everything: The Origin and Fate of the Universe

Chapter 1 - FIRST LECTURE - IDEAS ABOUT THE UNIVERSE

As long ago as 340 B.C. Aristotle, in his book On the Heavens, was able toput forward two good arguments for believing that the Earth was a roundball rather than a flat plate. First, he realized that eclipses of the moon werecaused by the Earth coming between the sun and the moon. The Earth's shad-ow on the moon was always round, which would be true only if the Earth wasspherical. If the Earth had been a flat disk, the shadow would have been elon-gated and elliptical, unless the eclipse always occurred at a time when the sunwas directly above the center of the disk.

Second, the Greeks knew from their travels that the Pole Star appeared lowerin the sky when viewed in the south than it did in more northerly regions.From the difference in the apparent position of the Pole Star in Egypt andGreece, Aristotle even quoted an estimate that the distance around the Earthwas four hundred thousand stadia. It is not known exactly what length a sta-dium was, but it may have been about two hundred yards. This would makeAristotle's estimate about twice the currently accepted figure.The Greeks even had a third argument that the Earth must be round, for whyelse does one first see the sails of a ship coming over the horizon and only latersee the hull? Aristotle thought that the Earth was stationary and that the sun,the moon, the planets, and the stars moved in circular orbits about the Earth.He believed this because he felt, for mystical reasons, that the Earth was thecenter of the universe and that circular motion was the most perfect.

This idea was elaborated by Ptolemy in the first century A.D. into a completecosmological model. The Earth stood at the center, surrounded by eightspheres, which carried the moon, the sun, the stars, and the five planets knownat the time: Mercury, Venus, Mars, Jupiter, and Saturn. The planets themselvesmoved on smaller circles attached to their respective spheres in order toaccount for their rather complicated observed paths in the sky. The outermostsphere carried the so-called fixed stars, which always stay in the same positionsrelative to each other but which rotate together across the sky. What laybeyond the last sphere was never made very clear, but it certainly was not partof mankind's observable universe.

Ptolemy's model provided a reasonably accurate system for

predicting thepositions of heavenly bodies in the sky. But in order to predict these positionscorrectly, Ptolemy had to make an assumption that the moon followed a paththat sometimes brought it twice as close to the Earth as at other times. Andthat meant that the moon had sometimes to appear twice as big as it usuallydoes. Ptolemy was aware of this flaw but nevertheless his model was generally,although not universally, accepted. It was adopted by the Christian church asthe picture of the universe that was in accordance with Scripture. It had thegreat advantage that it left lots of room outside the sphere of fixed stars forheaven and hell.

A much simpler model, however, was proposed in 1514 by a Polish priest,Nicholas Copernicus. At first, for fear of being accused of heresy, Copernicuspublished his model anonymously. His idea was that the sun was stationary atthe center and that the Earth and the planets moved in circular orbits aroundthe sun. Sadly for Copernicus, nearly a century passed before this idea was tobe taken seriously. Then two astronomers-the German, Johannes Kepler, andthe Italian, Galileo Galilei-started publicly to support the Copernican theo-ry, despite the fact that the orbits it predicted did not quite match the onesobserved. The death of the Aristotelian-Ptolemaic theory came in 1609. Inthat year Galileo started observing the night sky with a telescope, which hadjust been invented.

When he looked at the planet Jupiter, Galileo found that it was accompa-nied by several small satellites, or moons, which orbited around it. Thisimplied that everything did not have to orbit directly around the Earth asAristotle and Ptolemy had thought. It was, of course, still possible to believethat the Earth was stationary at the center of the universe, but that themoons of Jupiter moved on extremely complicated paths around the Earth,giving the appearance that they orbited Jupiter. However, Copernicus'stheory was much simpler.

At the same time, Kepler had modified Copernicus's theory, suggesting that theplanets moved not in circles, but in ellipses. The predictions now finallymatched the observations. As far as Kepler was concerned, elliptical orbits weremerely an ad hoc hypothesis-and a rather repugnant one at that becauseellipses were clearly less perfect than circles. Having discovered, almost by acci-dent, that elliptical orbits fitted the observations well, he could not reconcilewith his idea that the planets were made to orbit the sun by magnetic forces.An explanation was provided only much later, in 1687, when Newton pub-lished his Principia Mathematica Naturalis Causae. This was probably the

mostimportant single work ever published in the physical sciences. In it, Newtonnot only put forward a theory of how bodies moved in space and time, but healso developed the mathematics needed to analyze those motions. In addition,Newton postulated a law of universal gravitation. This said that each body inthe universe was attracted toward every other body by a force which wasstronger the more massive the bodies and the closer they were to each other.It was the same force which caused objects to fall to the ground. The story thatNewton was hit on the head by an apple is almost certainly apocryphal. AllNewton himself ever said was that the idea of gravity came to him as he sat ina contemplative mood, and was occasioned by the fall of an apple.

Newton went on to show that, according to his law, gravity causes the moonto move in an elliptical orbit around the Earth and causes the Earth and theplanets to follow elliptical paths around the sun. The Copernican model gotrid of Ptolemy's celestial spheres, and with them the idea that the universe hada natural boundary. The fixed stars did not appear to change their relative posi-tions as the Earth went around the sun. It therefore became natural to supposethat the fixed stars were objects like our sun but much farther away. This raiseda problem. Newton realized that, according to his theory of gravity, the starsshould attract each other; so, it seemed they could not remain essentiallymotionless. Would they not all fall together at some point?In a letter in 1691 to Richard Bentley, another leading thinker of his day,Newton argued that this would indeed happen if there were only a finite num-ber of stars. But he reasoned that if, on the other hand, there were an infinitenumber of stars distributed more or less uniformly over infinite space, thiswould not happen because there would not be any central point for them tofall to. This argument is an instance of the pitfalls that one can encounterwhen one talks about infinity.

In an infinite universe, every point can be regarded as the center because everypoint has an infinite number of stars on each side of it. The correct approach,it was realized only much later, is to consider the finite situation in which the

stars all fall in on each other. One then asks how things change if one addsmore stars roughly uniformly distributed outside this region. According toNewton's law, the extra stars would make no difference at all to the originalones, and so the stars would fall in just as fast. We can add as many stars as welike, but they will still always collapse in on themselves. We now know it isimpossible to have an infinite static model of the

universe in which gravity isalways attractive.

It is an interesting reflection on the general climate of thought before thetwentieth century that no one had suggested that the universe was expandingor contracting. It was generally accepted that either the universe had existedforever in an unchanging state or that it had been created at a finite time inthe past, more or less as we observe it today. In part, this may have been dueto people's tendency to believe in eternal truths as well as the comfort theyfound in the thought that even though they may grow old and die, the uni-verse is unchanging.

Even those who realized that Newton's theory of gravity showed that the uni-verse could not be static did not think to suggest that it might be expanding.Instead, they attempted to modify the theory by making the gravitational forcerepulsive at very large distances. This did not significantly affect their predic-tions of the motions of the planets. But it would allow an infinite distributionof stars to remain in equilibrium, with the attractive forces between nearbystars being balanced by the repulsive forces from those that were farther away.However, we now believe such an equilibrium would be unstable. If the starsin some region got only slightly near each other, the attractive forces betweenthem would become stronger and would dominate over the repulsive forces.This would mean that the stars would continue to fall toward each other. Onthe other hand, if the stars got a bit farther away from each other, the repul-sive forces would dominate and drive them farther apart.

Another objection to an infinite static universe is normally ascribed to theGerman philosopher Heinrich Olbers. In fact, various contemporaries ofNewton had raised the problem, and the Olbers article of 1823 was not eventhe first to contain plausible arguments on this subject. It was, however, thefirst to be widely noted. The difficulty is that in an infinite static universenearly every line or side would end on the surface of a star. Thus one wouldexpect that the whole sky would be as bright as the sun, even at night. Olbers'scounterargument was that the light from distant stars would be dimmed byabsorption by intervening matter. However, if that happened, the interveningmatter would eventually heat up until it glowed as brightly as the stars.

The only way of avoiding the conclusion that the whole of the night skyshould be as bright as the surface of the sun would be if the stars had not beenshining forever, but had turned on at some finite time in the past. In that case,the absorbing matter might not have heated up yet, or the light from distantstars might not yet

have reached us. And that brings us to the question of whatcould have caused the stars to have turned on in the first place.

THE BEGINNING OF THE UNIVERSE

The beginning of the universe had, of course, been discussed for a long time.According to a number of early cosmologies in the Jewish/Christian/Muslimtradition, the universe started at a finite and not very distant time in the past.One argument for such a beginning was the feeling that it was necessary tohave a first cause to explain the existence of the universe.Another argument was put forward by St. Augustine in his book, The City ofGod. He pointed out that civilization is progressing, and we remember whoperformed this deed or developed that technique. Thus man, and so also per-haps the universe, could not have been around all that long. For otherwise wewould have already progressed more than we have.

St. Augustine accepted a date of about 5000 B.C. for the creation of the uni-verse according to the book of Genesis. It is interesting that this is not so farfrom the end of the last Ice Age, about 10,000 B.C., which is when civilizationreally began. Aristotle and most of the other Greek philosophers, on the otherhand, did not like the idea of a creation because it made too much of divineintervention. They believed, therefore, that the human race and the worldaround it had existed, and would exist, forever. They had already consideredthe argument about progress, described earlier, and answered it by saying thatthere had been periodic floods or other disasters that repeatedly set the humanrace right back to the beginning of civilization.

When most people believed in an essentially static and unchanging universe,the question of whether or not it had a beginning was really one of meta-physics or theology. One could account for what was observed either way.Either the universe had existed forever, or it was set in motion at some finitetime in such a manner as to look as though it had existed forever. But in 1929,Edwin Hubble made the landmark observation that wherever you look, distantstars are moving rapidly away from us. In other words, the universe is expand-ing. This means that at earlier times objects would have been closer together.In fact, it seemed that there was a time about ten or twenty thousand millionyears ago when they were all at exactly the same place.

This discovery finally brought the question of the beginning of the universeinto the realm of science. Hubble's observations suggested that there was atime called the big bang when the universe was infinitesimally small and,therefore, infinitely dense. If there were events earlier than this time, thenthey could not

affect what happens at the present time. Their existence can beignored because it would have no observational consequences.

One may say that time had a beginning at the big bang, in the sense that ear-lier times simply could not be defined. It should be emphasized that this begin-ning in time is very different from those that had been considered previously.In an unchanging universe, a beginning in time is something that has to beimposed by some being outside the universe. There is no physical necessity fora beginning. One can imagine that God created the universe at literally anytime in the past. On the other hand, if the universe is expanding, there maybe physical reasons why there had to be a beginning. One could still believethat God created the universe at the instant of the big bang. He could evenhave created it at a later time in just such a way as to make it look as thoughthere had been a big bang. But it would be meaningless to suppose that it wascreated before the big bang. An expanding universe does not preclude a cre-ator, but it does place limits on when He might have carried out his job.

The Theory of Everything: The Origin and Fate of the Universe

Chapter 2 - SECOND LECTURE - THE EXPANDING UNIVERSE

Our sun and the nearby stars are all part of a vast collection of stars calledthe Milky Way galaxy. For a long time it was thought that this was thewhole universe. It was only in 1924 that the American astronomer EdwinHubble demonstrated that ours was not the only galaxy. There were, in fact,many others, with vast tracks of empty space between them. In order to provethis he needed to determine the distances to these other galaxies. We candetermine the distance of nearby stars by observing how they change positionas the Earth goes around the sun. But other galaxies are so far away that, unlikenearby stars, they really do appear fixed. Hubble was forced, therefore, to useindirect methods to measure the distances.

Now the apparent brightness of a star depends on two factors-luminosity andhow far it is from us. For nearby stars we can measure both their apparentbrightness and their distance, so we can work out their luminosity. Conversely,if we knew the luminosity of stars in other galaxies, we could work out theirdistance by measuring their apparent brightness. Hubble argued that therewere certain types of stars that always had the same luminosity when they werenear enough for us to measure. If, therefore, we found such stars in anothergalaxy, we could assume that they had the same luminosity. Thus, we couldcalculate the distance to that galaxy. If we could do this for a number of starsin the same galaxy, and our calculations always gave the same distance, wecould be fairly confident of our estimate. In this way, Edwin Hubble workedout the distances to nine different galaxies.

We now know that our galaxy is only one of some hundred thousand millionthat can be seen using modern telescopes, each galaxy itself containing somehundred thousand million stars. We live in a galaxy that is about one hundredthousand light-years across and is slowly rotating; the stars in its spiral armsorbit around its center about once every hundred million years. Our sun is justan ordinary, average-sized, yellow star, near the outer edge of one of the spiralarms. We have certainly come a long way since Aristotle and Ptolemy, whenwe thought that the Earth was the center of the universe.

Stars are so far away that they appear to us to be just pinpoints of light. Wecannot determine their size or shape. So

how can we tell different types of starsapart? For the vast majority of stars, there is only one correct characteristicfeature that we can observe-the color of their light. Newton discovered thatif light from the sun passes through a prism, it breaks up into its componentcolors-its spectrum-like in a rainbow. By focusing a telescope on anindividual star or galaxy, one can similarly observe the spectrum of the lightfrom that star or galaxy. Different stars have different spectra, but the relativebrightness of the different colors is always exactly what one would expect tofind in the light emitted by an object that is glowing red hot. This means thatwe can tell a star's temperature from the spectrum of its light. Moreover, wefind that certain very specific colors are missing from stars' spectra, and thesemissing colors may vary from star to star. We know that each chemical elementabsorbs the characteristic set of very specific colors. Thus, by matching each ofthose which are missing from a star's spectrum, we can determine exactlywhich elements are present in the star's atmosphere.

In the 1920s, when astronomers began to look at the spectra of stars in othergalaxies, they found something most peculiar: There were the same character-istic sets of missing colors as for stars in our own galaxy, but they were allshifted by the same relative amount toward the red end of the spectrum. Theonly reasonable explanation of this was that the galaxies were moving awayfrom us, and the frequency of the light waves from them was being reduced, orred-shifted, by the Doppler effect. Listen to a car passing on the road. As thecar is approaching, its engine sounds at a higher pitch, corresponding to ahigher frequency of sound waves; and when it passes and goes away, it soundsat a lower pitch. The behavior of light or radial waves is similar. Indeed, thepolice made use of the Doppler effect to measure the speed of cars by measur-ing the frequency of pulses of radio waves reflected off them.

In the years following his proof of the existence of other galaxies, Hubble spenthis time cataloging their distances and observing their spectra. At that timemost people expected the galaxies to be moving around quite randomly, and soexpected to find as many spectra which were blue-shifted as ones which werered-shifted. It was quite a surprise, therefore, to find that the galaxies allappeared red-shifted. Every single one was moving away from us. More surpris-ing still was the result which Hubble published in 1929: Even the size of thegalaxy's red shift was not random, but was directly proportional to the galaxy'sdistance from us. Or, in other words, the farther a galaxy was, the faster it

wasmoving away. And that meant that the universe could not be static, as every-one previously thought, but was in fact expanding. The distance between thedifferent galaxies was growing all the time.

The discovery that the universe was expanding was one of the great intellec-tual revolutions of the twentieth century. With hindsight, it is easy to wonderwhy no one had thought of it before. Newton and others should have realizedthat a static universe would soon start to contract under the influence ofgravity. But suppose that, instead of being static, the universe was expanding.If it was expanding fairly slowly, the force of gravity would cause it eventuallyto stop expanding and then to start contracting. However, if it was expandingat more than a certain critical rate, gravity would never be strong enough tostop it, and the universe would continue to expand forever. This is a bit likewhat happens when one fires a rocket upward from the surface of the Earth. Ifit has a fairly low speed, gravity will eventually stop the rocket and it will startfalling back. On the other hand, if the rocket has more than a certain criticalspeed-about seven miles a second-gravity will not be strong enough to pull itback, so it will keep going away from the Earth forever.

This behavior of the universe could have been predicted from Newton's theoryof gravity at any time in the nineteenth, the eighteenth, or even the late sev-enteenth centuries. Yet so strong was the belief in a static universe that it per-sisted into the early twentieth century. Even when Einstein formulated thegeneral theory of relativity in 1915, he was sure that the universe had to bestatic. He therefore modified his theory to make this possible, introducing a so-called cosmological constant into his equations. This was a new "antigravity"force, which, unlike other forces, did not come from any particular source, butwas built into the very fabric of space-time. His cosmological constant gavespace-time an inbuilt tendency to expand, and this could be made to exactlybalance the attraction of all the matter in the universe so that a static universewould result.

Only one man, it seems, was willing to take general relativity at face value.While Einstein and other physicists were looking for ways of avoiding generalrelativity's prediction of a nonstatic universe, the Russian physicist AlexanderFriedmann instead set about explaining it.

THE FRIEDMANN MODELSThe equations of general relativity, which determined how the universeevolves in time, are too complicated to solve in detail. So what Friedmanndid, instead, was to make two very simple assumptions about the universe:that

the universe looks identical in whichever direction we look, and thatthis would also be true if we were observing the universe from anywhere else.On the basis of general relativity and these two assumptions, Friedmannshowed that we should not expect the universe to be static. In fact, in 1922,several years before Edwin Hubble's discovery, Friedmann predicted exactlywhat Hubble found.

The assumption that the universe looks the same in every direction is clearly nottrue in reality. For example, the other stars in our galaxy form a distinct band oflight across the night sky called the Milky Way. But if we look at distant galax-ies, there seems to be more or less the same number of them in each direction.So the universe does seem to be roughly the same in every direction, providedone views it on a large scale compared to the distance between galaxies.For a long time this was sufficient justification for Friedmann's assumption-as a rough approximation to the real universe. But more recently a lucky acci-dent uncovered the fact that Friedmann's assumption is in fact a remarkablyaccurate description of our universe. In 1965, two American physicists, ArnoPenzias and Robert Wilson, were working at the Bell Labs in New Jersey onthe design of a very sensitive microwave detector for communicating withorbiting satellites. They were worried when they found that their detector waspicking up more noise than it ought to, and that the noise did not appear tobe coming from any particular direction. First they looked for bird droppingson their detector and checked for other possible malfunctions, but soon ruledthese out. They knew that any noise from within the atmosphere would bestronger when the detector is not pointing straight up than when it is, becausethe atmosphere appears thicker when looking at an angle to the vertical.The extra noise was the same whichever direction the detector pointed, so itmust have come from outside the atmosphere. It was also the same day andnight throughout the year, even though the Earth was rotating on its axis andorbiting around the sun. This showed that the radiation must come frombeyond the solar system, and even from beyond the galaxy, as otherwise itwould vary as the Earth pointed the detector in different directions.In fact, we know that the radiation must have traveled to us across most ofthe observable universe. Since it appears to be the same in different direc-tions, the universe must also be the same in every direction, at least on a largescale. We now know that whichever direction we look in, this noise nevervaries by more than one part in ten thousand. So Penzias and Wilson hadunwittingly stumbled across a remarkably

accurate confirmation ofFriedmann's first assumption.

At roughly the same time, two American physicists at nearby PrincetonUniversity, Bob Dicke and Jim Peebles, were also taking an interest inmicrowaves. They were working on a suggestion made by George Gamow,once a student of Alexander Friedmann, that the early universe should havebeen very hot and dense, glowing white hot. Dicke and Peebles argued that weshould still be able to see this glowing, because light from very distant partsof the early universe would only just be reaching us now. However, theexpansion of the universe meant that this light should be so greatly red-shift-ed that it would appear to us now as microwave radiation. Dicke and Peebleswere looking for this radiation when Penzias and Wilson heard about theirwork and realized that they had already found it. For this, Penzias andWilson were awarded the Nobel Prize in 1978, which seems a bit hard onDicke and Peebles.

Now at first sight, all this evidence that the universe looks the same whichev-er direction we look in might seem to suggest there is something special aboutour place in the universe. In particular, it might seem that if we observe allother galaxies to be moving away from us, then we must be at the center of theuniverse. There is, however, an alternative explanation: The universe mightalso look the same in every direction as seen from any other galaxy. This, as wehave seen, was Friedmann's second assumption.

We have no scientific evidence for or against this assumption. We believe itonly on grounds of modesty. It would be most remarkable if the universelooked the same in every direction around us, but not around other points inthe universe. In Friedmann's model, all the galaxies are moving directly awayfrom each other. The situation is rather like steadily blowing up a balloonwhich has a number of spots painted on it. As the balloon expands, the dis-tance between any two spots increases, but there is no spot that can be said tobe the center of the expansion. Moreover, the farther apart the spots are, thefaster they will be moving apart. Similarly, in Friedmann's model the speed atwhich any two galaxies are moving apart is proportional to the distancebetween them. So it predicted that the red shift of a galaxy should be directlyproportional to its distance from us, exactly as Hubble found.

Despite the success of his model and his prediction of Hubble's observations,Friedmann's work remained largely unknown in the West. It became knownonly after similar models were discovered in 1935 by the American physicistHoward Robertson and the

British mathematician Arthur Walker, in responseto Hubble's discovery of the uniform expansion of the universe.

Although Friedmann found only one, there are in fact three different kinds ofmodels that obey Friedmann's two fundamental assumptions. In the firstkind-which Friedmann found-the universe is expanding so sufficientlyslowly that the gravitational attraction between the different galaxies causesthe expansion to slow down and eventually to stop. The galaxies then start tomove toward each other and the universe contracts. The distance between twoneighboring galaxies starts at zero, increases to a maximum, and then decreasesback down to zero again.

In the second kind of solution, the universe is expanding so rapidly that thegravitational attraction can never stop it, though it does slow it down a bit.The separation between neighboring galaxies in this model starts at zero, andeventually the galaxies are moving apart at a steady speed.

Finally, there is a third kind of solution, in which the universe is expandingonly just fast enough to avoid recollapse. In this case the separation also startsat zero, and increases forever. However, the speed at which the galaxies aremoving apart gets smaller and smaller, although it never quite reaches zero.

A remarkable feature of the first kind of Friedmann model is that the universeis not infinite in space, but neither does space have any boundary. Gravity isso strong that space is bent round onto itself, making it rather like the surfaceof the Earth. If one keeps traveling in a certain direction on the surface of theEarth, one never comes up against an impassable barrier or falls over the edge,but eventually comes back to where one started. Space, in the first Friedmannmodel, is just like this, but with three dimensions instead of two for the Earth'ssurface. The fourth dimension-time-is also finite in extent, but it is like aline with two ends or boundaries, a beginning and an end. We shall see laterthat when one combines general relativity with the uncertainty principle ofquantum mechanics, it is possible for both space and time to be finite withoutany edges or boundaries. The idea that one could go right around the universeand end up where one started makes good science fiction, but it doesn't havemuch practical significance because it can be shown that the universe wouldrecollapse to zero size before one could get round. You would need to travelfaster than light in order to end up where you started before the universe cameto an end-and that is not allowed.

But which Friedmann model describes our universe? Will the universe eventu-ally stop expanding and start contracting, or will

it expand forever? To answerthis question we need to know the present rate of expansion of the universeand its present average density. If the density is less than a certain criticalvalue, determined by the rate of expansion, the gravitational attraction will betoo weak to halt the expansion. If the density is greater than the critical value,gravity will stop the expansion at some time in the future and cause theuniverse to recollapse.

We can determine the present rate of expansion by measuring the velocities atwhich other galaxies are moving away from us, using the Doppler effect. Thiscan be done very accurately. However, the distances to the galaxies are notvery well known because we can only measure them indirectly. So all we knowis that the universe is expanding by between 5 percent and 10 percent everythousand million years. However, our uncertainty about the present averagedensity of the universe is even greater.

If we add up the masses of all the stars that we can see in our galaxy and othergalaxies, the total is less than one-hundredth of the amount required to haltthe expansion of the universe, even in the lowest estimate of the rate of expan-sion. But we know that our galaxy and other galaxies must contain a largeamount of dark matter which we cannot see directly, but which we know mustbe there because of the influence of its gravitational attraction on the orbits ofstars and gas in the galaxies. Moreover, most galaxies are found in clusters, andwe can similarly infer the presence of yet more dark matter in between thegalaxies in these clusters by its effect on the motion of the galaxies. When weadd up all this dark matter, we still get only about one-tenth of the amountrequired to halt the expansion. However, there might be some other form ofmatter which we have not yet detected and which might still raise the averagedensity of the universe up to the critical value needed to halt the expansion.The present evidence, therefore, suggests that the universe will probablyexpand forever. But don't bank on it. All we can really be sure of is that evenif the universe is going to recollapse, it won't do so for at least another tenthousand million years, since it has already been expanding for at least thatlong. This should not unduly worry us since by that time, unless we havecolonies beyond the solar system, mankind will long since have died out,extinguished along with the death of our sun.

THE BIG BANG

All of the Friedmann solutions have the feature that at some time in thepast, between ten and twenty thousand million years ago, the distancebetween neighboring galaxies must have been zero. At that time, which wecall the big bang, the density of the

universe and the curvature of space-timewould have been infinite. This means that the general theory of relativity-on which Friedmann's solutions are based-predicts that there is a singularpoint in the universe.

All our theories of science are formulated on the assumption that space-timeis smooth and nearly flat, so they would all break down at the big bang singu-larity, where the curvature of space-time is infinite. This means that even ifthere were events before the big bang, one could not use them to determinewhat would happen afterward, because predictability would break down at thebig bang. Correspondingly, if we know only what has happened since the bigbang, we could not determine what happened beforehand. As far as we areconcerned, events before the big bang can have no consequences, so theyshould not form part of a scientific model of the universe. We should thereforecut them out of the model and say that time had a beginning at the big bang.Many people do not like the idea that time has a beginning, probably becauseit smacks of divine intervention. (The Catholic church, on the other hand, hadseized on the big bang model and in 1951 officially pronounced it to be inaccordance with the Bible.) There were a number of attempts to avoid the con-clusion that there had been a big bang. The proposal that gained widest supportwas called the steady state theory. It was suggested in 1948 by two refugees fromNazi-occupied Austria, Hermann Bondi and Thomas Gold, together with theBriton Fred Hoyle, who had worked with them on the development of radarduring the war. The idea was that as the galaxies moved away from each other,new galaxies were continually forming in the gaps in between, from newmatter that was being continually created. The universe would therefore lookroughly the same at all times as well as at all points of space.

The steady state theory required a modification of general relativity to allowfor the continual creation of matter, but the rate that was involved was solow-about one particle per cubic kilometer per year-that it was not in con-flict with experiment. The theory was a good scientific theory, in the sensethat it was simple and it made definite predictions that could be tested byobservation. One of these predictions was that the number of galaxies or sim-ilar objects in any given volume of space should be the same wherever andwhenever we look in the universe.

In the late 1950s and early 1960s, a survey of sources of radio waves from outerspace was carried out at Cambridge by a group of astronomers led by MartinRyle. The Cambridge group showed that most of these radio sources must lieoutside our galaxy, and

also that there were many more weak sources thanstrong ones. They interpreted the weak sources as being the more distant ones,and the stronger ones as being near. Then there appeared to be fewer sourcesper unit volume of space for the nearby sources than for the distant ones.

This could have meant that we were at the center of a great region in the uni-verse in which the sources were fewer than elsewhere. Alternatively, it couldhave meant that the sources were more numerous in the past, at the time thatthe radio waves left on their journey to us, than they are now. Either explana-tion contradicted the predictions of the steady state theory. Moreover, thediscovery of the microwave radiation by Penzias and Wilson in 1965 also indi-cated that the universe must have been much denser in the past. The steadystate theory therefore had regretfully to be abandoned.

Another attempt to avoid the conclusion that there must have been a big bangand, therefore, a beginning of time, was made by two Russian scientists,Evgenii Lifshitz and Isaac Khalatnikov, in 1963. They suggested that the bigbang might be a peculiarity of Friedmann's models alone, which after all wereonly approximations to the real universe. Perhaps, of all the models that wereroughly like the real universe, only Friedmann's would contain a big bang sin-gularity. In Friedmann's models, the galaxies are all moving directly away fromeach other. So it is not surprising that at some time in the past they were all atthe same place. In the real universe, however, the galaxies are not just movingdirectly away from each other-they also have small sideways velocities. So inreality they need never have been all at exactly the same place, only very closetogether. Perhaps, then, the current expanding universe resulted not from a bigbang singularity, but from an earlier contracting phase; as the universe had col-lapsed, the particles in it might not have all collided, but they might haveflown past and then away from each other, producing the present expansion ofthe universe. How then could we tell whether the real universe should havestarted out with a big bang?

What Lifshitz and Khalatnikov did was to study models of the universe whichwere roughly like Friedmann's models but which took account of the irregular-ities and random velocities of galaxies in the real universe. They showed thatsuch models could start with a big bang, even though the galaxies were nolonger always moving directly away from each other. But they claimed thatthis was still only possible in certain exceptional models in which the galaxieswere all moving in just the right way. They

argued that since there seemed tobe infinitely more Friedmann-like models without a big bang singularity thanthere were with one, we should conclude that it was very unlikely that therehad been a big bang. They later realized, however, that there was a much moregeneral class of Friedmann-like models which did have singularities, and inwhich the galaxies did not have to be moving in any special way. They there-fore withdrew their claim in 1970.

The work of Lifshitz and Khalatnikov was valuable because it showed that theuniverse could have had a singularity-a big bang-if the general theory of rel-ativity was correct. However, it did not resolve the crucial question: Does gen-eral relativity predict that our universe should have the big bang, a beginningof time? The answer to this came out of a completely different approach start-ed by a British physicist, Roger Penrose, in 1965. He used the way light conesbehave in general relativity, and the fact that gravity is always attractive, toshow that a star that collapses under its own gravity is trapped in a region whoseboundary eventually shrinks to zero size. This means that all the matter in thestar will be compressed into a region of zero volume, so the density of matterand the curvature of space-time become infinite. In other words, one has a sin-gularity contained within a region of space-time known as a black hole.

At first sight, Penrose's result didn't have anything to say about the questionof whether there was a big bang singularity in the past. However, at the timethat Penrose produced his theorem, I was a research student desperately look-ing for a problem with which to complete my Ph.D. thesis. I realized that if onereversed the direction of time in Penrose's theorem so that the collapse becamean expansion, the conditions of his theorem would still hold, provided theuniverse were roughly like a Friedmann model on large scales at the presenttime. Penrose's theorem had shown that any collapsing star must end in asingularity; the time-reversed argument showed that any Friedmann-likeexpanding universe must have begun with a singularity. For technical reasons,Penrose's theorem required that the universe be infinite in space. So I coulduse it to prove that there should be a singularity only if the universe wasexpanding fast enough to avoid collapsing again, because only that Friedmannmodel was infinite in space.

During the next few years I developed new mathematical techniques toremove this and other technical conditions from the theorems that provedthat singularities must occur. The final result was a joint paper by Penroseand myself in 1970, which proved that there must have been a big bang singu-larity provided only

that general relativity is correct and that the universecontains as much matter as we observe.

There was a lot of opposition to our work, partly from the Russians, whofollowed the party line laid down by Lifshitz and Khalatnikov, and partly frompeople who felt that the whole idea of singularities was repugnant and spoiledthe beauty of Einstein's theory. However, one cannot really argue with themathematical theorem. So it is now generally accepted that the universe musthave a beginning.

The Theory of Everything: The Origin and Fate of the Universe

Chapter 3 - THIRD LECTURE - BLACK HOLES

The term black hole is of very recent origin. It was coined in 1969 by theAmerican scientist John Wheeler as a graphic description of an idea thatgoes back at least two hundred years. At that time there were two theoriesabout light. One was that it was composed of particles; the other was that itwas made of waves. We now know that really both theories are correct. By thewave/particle duality of quantum mechanics, light can be regarded as both awave and a particle. Under the theory that light was made up of waves, it wasnot clear how it would respond to gravity. But if light were composed of parti-cles, one might expect them to be affected by gravity in the same way thatcannonballs, rockets, and planets are.

On this assumption, a Cambridge don, John Michell, wrote a paper in 1783in the Philosophical Transactions of the Royal Society of London. In it, he point-ed out that a star that was sufficiently massive and compact would have sucha strong gravitational field that light could not escape. Any light emittedfrom the surface of the star would be dragged back by the star's gravitationalattraction before it could get very far. Michell suggested that there might bea large number of stars like this. Although we would not be able to see thembecause the light from them would not reach us, we would still feel their grav-itational attraction. Such objects are what we now call black holes, becausethat is what they are-black voids in space.

A similar suggestion was made a few years later by the French scientist theMarquis de Laplace, apparently independently of Michell. Interestinglyenough, he included it in only the first and second editions of his book, TheSystem of the World, and left it out of later editions; perhaps he decided that itwas a crazy idea. In fact, it is not really consistent to treat light like cannon-balls in Newton's theory of gravity because the speed of light is fixed. A can-nonball fired upward from the Earth will be slowed down by gravity and willeventually stop and fall back. A photon, however, must continue upward at aconstant speed. How, then, can Newtonian gravity affect light? A consistenttheory of how gravity affects light did not come until Einstein proposed gen-eral relativity in 1915; and even then it was a long time before the implica-tions of the theory for massive stars were worked out.

To understand how a black hole might be formed, we first need an understand-ing of the life cycle of a star. A star is formed when a large amount of gas, most-ly hydrogen, starts to collapse

in on itself due to its gravitational attraction. Asit contracts, the atoms of the gas collide with each other more and more frequently and at greater and greater speeds-the gas heats up. Eventually the gaswill be so hot that when the hydrogen atoms collide they no longer bounce offeach other but instead merge with each other to form helium atoms. The heatreleased in this reaction, which is like a controlled hydrogen bomb, is whatmakes the stars shine. This additional heat also increases the pressure of thegas until it is sufficient to balance the gravitational attraction, and the gasstops contracting. It is a bit like a balloon where there is a balance between thepressure of the air inside, which is trying to make the balloon expand, and thetension in the rubber, which is trying to make the balloon smaller.

The stars will remain stable like this for a long time, with the heat from thenuclear reactions balancing the gravitational attraction. Eventually, however,the star will run out of its hydrogen and other nuclear fuels. And paradoxical-ly, the more fuel a star starts off with, the sooner it runs out. This is becausethe more massive the star is, the hotter it needs to be to balance its gravita-tional attraction. And the hotter it is, the faster it will use up its fuel. Our sunhas probably got enough fuel for another five thousand million years or so, butmore massive stars can use up their fuel in as little as one hundred millionyears, much less than the age of the universe. When the star runs out of fuel,it will start to cool off and so to contract. What might happen to it then wasonly first understood at the end of the 1920s.

In 1928 an Indian graduate student named Subrahmanyan Chandrasekhar setsail for England to study at Cambridge with the British astronomer Sir ArthurEddington. Eddington was an expert on general relativity. There is a story thata journalist told Eddington in the early 1920s that he had heard there wereonly three people in the world who understood general relativity. Eddingtonreplied, “I am trying to think who the third person is.”During his voyage from India, Chandrasekhar worked out how big a star couldbe and still separate itself against its own gravity after it had used up all itsfuel. The idea was this: When the star becomes small, the matter particles getvery near each other. But the Pauli exclusion principle says that two matterparticles cannot have both the same position and the same velocity. The mat-ter particles must therefore have very different velocities. This makes themmove away from each other, and so tends to make the star expand. A star cantherefore maintain itself at a constant radius by a balance between the attrac-tion of gravity and the repulsion that arises from the exclusion principle, justas earlier in its life

the gravity was balanced by the heat.

Chandrasekhar realized, however, that there is a limit to the repulsion that theexclusion principle can provide. The theory of relativity limits the maximumdifference in the velocities of the matter particles in the star to the speed oflight. This meant that when the star got sufficiently dense, the repulsioncaused by the exclusion principle would be less than the attraction of gravity.Chandrasekhar calculated that a cold star of more than about one and a halftimes the mass of the sun would not be able to support itself against its owngravity. This mass is now known as the Chandrasekhar limit.

This had serious implications for the ultimate fate of massive stars. If a star'smass is less than the Chandrasekhar limit, it can eventually stop contractingand settle down to a possible final state as a white dwarf with a radius of a fewthousand miles and a density of hundreds of tons per cubic inch. A white dwarfis supported by the exclusion principle repulsion between the electrons in itsmatter. We observe a large number of these white dwarf stars. One of the firstto be discovered is the star that is orbiting around Sirius, the brightest star inthe night sky.

It was also realized that there was another possible final state for a star alsowith a limiting mass of about one or two times the mass of the sun, but muchsmaller than even the white dwarf. These stars would be supported by theexclusion principle repulsion between the neutrons and protons, rather thanbetween the electrons. They were therefore called neutron stars. They wouldhave had a radius of only ten miles or so and a density of hundreds of millionsof tons per cubic inch. At the time they were first predicted, there was no waythat neutron stars could have been observed, and they were not detected untilmuch later.

Stars with masses above the Chandrasekhar limit, on the other hand, have abig problem when they come to the end of their fuel. In some cases they mayexplode or manage to throw off enough matter to reduce their mass below thelimit, but it was difficult to believe that this always happened, no matter howbig the star. How would it know that it had to lose weight? And even if everystar managed to lose enough mass, what would happen if you added more massto a white dwarf or neutron star to take it over the limit? Would it collapse toinfinite density?

Eddington was shocked by the implications of this and refused to believeChandrasekhar's result. He thought it was simply not possible that a star couldcollapse to a point. This was the view of most scientists. Einstein himself wrotea paper in which he claimed that stars would not shrink to zero size.The hos-tility of

other scientists, particularly of Eddington, his former teacher and theleading authority on the structure of stars, persuaded Chandrasekhar to aban-don this line of work and turn instead to other problems in astronomy.However, when he was awarded the Nobel Prize in 1983, it was, at least inpart, for his early work on the limiting mass of cold stars.

Chandrasekhar had shown that the exclusion principle could not halt the col-lapse of a star more massive than the Chandrasekhar limit. But the problem ofunderstanding what would happen to such a star, according to general relativ-ity, was not solved until 1939 by a young American, Robert Oppenheimer. Hisresult, however, suggested that there would be no observational consequencesthat could be detected by the telescopes of the day. Then the war intervenedand Oppenheimer himself became closely involved in the atom bomb project.And after the war the problem of gravitational collapse was largely forgottenas most scientists were then interested in what happens on the scale of theatom and its nucleus. In the 1960s, however, interest in the large-scale prob-lems of astronomy and cosmology was revived by a great increase in the num-ber and range of astronomical observations brought about by the applicationof modern technology. Oppenheimer's work was then rediscovered andextended by a number of people.

The picture that we now have from Oppenheimer's work is as follows: Thegravitational field of the star changes the paths of light rays in space-time fromwhat they would have been had the star not been present. The light cones,which indicate the paths followed in space and time by flashes of light emit-ted from their tips, are bent slightly inward near the surface of the star. Thiscan be seen in the bending of light from distant stars that is observed duringan eclipse of the sun. As the star contracts, the gravitational field at its surfacegets stronger and the light cones get bent inward more. This makes it moredifficult for light from the star to escape, and the light appears dimmer andredder to an observer at a distance.

Eventually, when the star has shrunk to a certain critical radius, the gravita-tional field at the surface becomes so strong that the light cones are bentinward so much that the light can no longer escape. According to the theoryof relativity, nothing can travel faster than light. Thus, if light cannot escape,neither can anything else. Everything is dragged back by the gravitationalfield. So one has a set of events, a region of space-time, from which it is notpossible to escape to reach a distant observer. This region is what we now calla black hole. Its

boundary is called the event horizon. It coincides with thepaths of the light rays that just fail to escape from the black hole.

In order to understand what you would see if you were watching a star collapseto form a black hole, one has to remember that in the theory of relativity thereis no absolute time. Each observer has his own measure of time. The time forsomeone on a star will be different from that for someone at a distance, becauseof the gravitational field of the star. This effect has been measured in an exper-iment on Earth with clocks at the top and bottom of a water tower. Supposean intrepid astronaut on the surface of the collapsing star sent a signal everysecond, according to his watch, to his spaceship orbiting about the star. Atsome time on his watch, say eleven o'clock, the star would shrink below thecritical radius at which the gravitational field became so strong that the signalswould no longer reach the spaceship.

His companions watching from the spaceship would find the intervals betweensuccessive signals from the astronaut getting longer and longer as eleveno'clock approached. However, the effect would be very small before 10:59:59.They would have to wait only very slightly more than a second between theastronaut's 10:59:58 signal and the one that he sent when his watch read10:59:59, but they would have to wait forever for the eleven o'clock signal.The light waves emitted from the surface of the star between 10:59:59 andeleven o'clock, by the astronaut's watch, would be spread out over an infiniteperiod of time, as seen from the spaceship.

The time interval between the arrival of successive waves at the spaceshipwould get longer and longer, and so the light from the star would appearredder and redder and fainter and fainter. Eventually the star would be so dimthat it could no longer be seen from the spaceship. All that would be left wouldbe a black hole in space. The star would, however, continue to exert the samegravitational force on the spaceship. This is because the star is still visible tothe spaceship, at least in principle. It is just that the light from the surface isso red-shifted by the gravitational field of the star that it cannot be seen.However, the red shift does not affect the gravitational field of the star itself.Thus, the spaceship would continue to orbit the black hole.

The work that Roger Penrose and I did between 1965 and 1970 showed that,according to general relativity, there must be a singularity of infinite densitywithin the black hole. This is rather like the big bang at the beginning of time,only it would be an end of time for the collapsing body and the astronaut. Atthe singularity, the laws of science and our ability to predict the

future wouldbreak down. However, any observer who remained outside the black holewould not be affected by this failure of predictability, because neither light norany other signal can reach them from the singularity.

This remarkable fact led Roger Penrose to propose the cosmic censorshiphypothesis, which might be paraphrased as “God abhors a naked singularity.”In other words, the singularities produced by gravitational collapse occur onlyin places like black holes, where they are decently hidden from outside viewby an event horizon. Strictly, this is what is known as the weak cosmic censor-ship hypothesis: protect obervers who remain outside the black hole from theconsequences of the breakdown of predictability that occurs at the singularity.But it does nothing at all for the poor unfortunate astronaut who falls into thehole. Shouldn’t God protect his modesty as well?

There are some solutions of the equations of general relativity in which it ispossible for our astronaut to see a naked singularity. He may be able to avoidhitting the singularity and instead fall through a “worm hole” and come out inanother region of the universe. This would offer great possibilities for travel inspace and time, but unfortunately it seems that the solutions may all be high-ly unstable. The least disturbance, such as the presence of an astronaut, maychange them so that the astronaut cannot see the singularity until he hits itand his time comes to an end. In other words, the singularity always lies in hisfuture and never in his past.

The strong version of the cosmic censorship hypothesis states that in a realis-tic solution, the singularities always lie either entirely in the future, like thesingularities of gravitational collapse, or entirely in the past, like the big bang.It is greatly to be hoped that some version of the censorship hypothesis holds,because close to naked singularities it may be possible to travel into the past.While this would be fine for writers of science fiction, it would mean that noone’s life would ever be safe. Someone might go into the past and kill yourfather or mother before you were conceived.

In a gravitational collapse to form a black hole, the movements would bedammed by the emission of gravitational waves. One would therefore expectthat it would not be too long before the black hole would settle down to a sta-tionary state. It was generally supposed that this final stationary state wouldde pend on the details of the body that had collapsed to form the black hole.The black hole might have any shape or size, and its shape might not even befixed, but instead be

pulsating.However, in 1967, the study of black holes was revolutionized by a paper writ-ten in Dublin by Werner Israel. Israel showed that any black hole that is notrotating must be perfectly round or spherical. Its size, moreover, would dependonly on its mass. It could, in fact, be described by a particular solution ofEinstein's equations that had been known since 1917, when it had been foundby Karl Schwarzschild shortly after the discovery of general relativity. At first,Israel's result was interpreted by many people, including Israel himself, as evi-dence that black holes would form only from the collapse of bodies that wereperfectly round or spherical. As no real body would be perfectly spherical, thismeant that, in general, gravitational collapse would lead to naked singularities.There was, however, a different interpretation of Israel's result, which wasadvocated by Roger Penrose and John Wheeler in particular. This was that ablack hole should behave like a ball of fluid. Although a body might start offin an unspherical state, as it collapsed to form a black hole it would settle downto a spherical state due to the emission of gravitational waves. Further calcu-lations supported this view and it came to be adopted generally.

Israel's result had dealt only with the case of black holes formed from nonro-tating bodies. On the analogy with a ball of fluid, one would expect that ablack hole made by the collapse of a rotating body would not be perfectlyround. It would have a bulge round the equator caused by the effect of the rota-tion. We observe a small bulge like this in the sun, caused by its rotation onceevery twenty-five days or so. In 1963, Roy Kerr, a New Zealander, had found aset of black-hole solutions of the equations of general relativity more generalthan the Schwarzschild solutions. These "Kerr" black holes rotate at aconstant rate, their size and shape depending only on their mass and rate ofrotation. If the rotation was zero, the black hole was perfectly round and thesolution was identical to the Schwarzschild solution. But if the rotation wasnonzero, the black hole bulged outward near its equator. It was therefore nat-ural to conjecture that a rotating body collapsing to form a black hole wouldend up in a state described by the Kerr solution.

In 1970, a colleague and fellow research student of mine, Brandon Carter, tookthe first step toward proving this conjecture. He showed that, provided a sta-tionary rotating black hole had an axis of symmetry, like a spinning top, its sizeand shape would depend only on its mass and rate of rotation. Then, in 1971,I proved that any stationary rotating black hole would indeed have such anaxis of symmetry. Finally, in 1973, David Robinson at

Kings College, London,used Carter's and my results to show that the conjecture had been correct:Such a black hole had indeed to be the Kerr solution.

So after gravitational collapse a black hole must settle down into a state inwhich it could be rotating, but not pulsating. Moreover, its size and shapewould depend only on its mass and rate of rotation, and not on the nature ofthe body that had collapsed to form it. This result became known by themaxim "A black hole has no hair." It means that a very large amount of infor-mation about the body that has collapsed must be lost when a black hole isformed, because afterward all we can possibly measure about the body is itsmass and rate of rotation. The significance of this will be seen in the next lec-ture. The no-hair theorem is also of great practical importance because it sogreatly restricts the possible types of black holes. One can therefore makedetailed models of objects that might contain black holes, and compare thepredictions of the models with observations.

Black holes are one of only a fairly small number of cases in the history of sci-ence where a theory was developed in great detail as a mathematical modelbefore there was any evidence from observations that it was correct. Indeed,this used to be the main argument of opponents of black holes. How could onebelieve in objects for which the only evidence was calculations based on thedubious theory of general relativity?

In 1963, however, Maarten Schmidt, an astronomer at the Mount PalomarObservatory in California, found a faint, starlike object in the direction of thesource of radio waves called 3C273-that is, source number 273 in the thirdCambridge catalog of radio sources. When he measured the red shift of theobject, he found it was too large to be caused by a gravitational field: If it hadbeen a gravitational red shift, the object would have to be so massive and sonear to us that it would disturb the orbits of planets in the solar system. Thissuggested that the red shift was instead caused by the expansion of the uni-verse, which in turn meant that the object was a very long way away. And tobe visible at such a great distance, the object must be very bright and must beemitting a huge amount of energy.

The only mechanism people could think of that would produce such largequantities of energy seemed to be the gravitational collapse not just of a starbut of the whole central region of a galaxy. A number of other similar "quasi-stellar objects," or quasars, have since been discovered, all with large red shifts.But they are all too far away, and too difficult, to observe to provide conclu-sive evidence of black holes.

Further encouragement for the existence of black holes came in 1967 with thediscovery by a research student at Cambridge, Jocelyn Bell, of some objects inthe sky that were emitting regular pulses of radio waves. At first, Jocelyn andher supervisor, Anthony Hewish, thought that maybe they had made contactwith an alien civilization in the galaxy. Indeed, at the seminar at which theyannounced their discovery, I remember that they called the first four sourcesto be found LGM 1-4, LGM standing for "Little Green Men."

In the end, however, they and everyone else came to the less romantic conclu-sion that these objects, which were given the name pulsars, were in fact justrotating neutron stars. They were emitting pulses of radio waves because of acomplicated indirection between their magnetic fields and surrounding matter.This was bad news for writers of space westerns, but very hopeful for the smallnumber of us who believed in black holes at that time. It was the first positiveevidence that neutron stars existed. A neutron star has a radius of about tenmiles, only a few times the critical radius at which a star becomes a black hole.If a star could collapse to such a small size, it was not unreasonable to expectthat other stars could collapse to even smaller size and become black holes.How could we hope to detect a black hole, as by its very definition it does notemit any light? It might seem a bit like looking for a black cat in a coal cellar.Fortunately, there is a way, since as John Michell pointed out in his pioneer-ing paper in 1783, a black hole still exerts a gravitational force on nearbyobjects. Astronomers have observed a number of systems in which two starsorbit around each other, attracted toward each other by gravity. They alsoobserved systems in which there is only one visible star that is orbiting aroundsome unseen companion.

One cannot, of course, immediately conclude that the companion is a blackhole. It might merely be a star that is too faint to be seen. However, some ofthese systems, like the one called Cygnus X-I, are also strong sources of X rays.The best explanation for this phenomenon is that the X rays are generated bymatter that has been blown off the surface of the visible star. As it falls towardthe unseen companion, it develops a spiral motion-rather like water runningout of a bath-and it gets very hot, emitting X rays. For this mechanism towork, the unseen object has to be very small, like a white dwarf, neutron star,or black hole.

Now, from the observed motion of the visible star, one can determine the low-est possible mass of the unseen object. In the

case of Cygnus X-I, this is aboutsix times the mass of the sun. According to Chandrasekhar's result, this is toomuch for the unseen object to be a white dwarf. It is also too large a mass tobe a neutron star. It seems, therefore, that it must be a black hole.There are other models to explain Cygnus X-I that do not include a blackhole, but they are all rather far-fetched. A black hole seems to be the onlyreally natural explanation of the observations. Despite this, I have a bet withKip Thorne of the California Institute of Technology that in fact Cygnus X-Idoes not contain a black hole. This is a form of insurance policy for me. I havedone a lot of work on black holes, and it would all be wasted if it turned outthat black holes do not exist. But in that case, I would have the consolation ofwinning my bet, which would bring me four years of the magazine Private Eye.

If black holes do exist, Kip will get only one year of Penthouse, because whenwe made the bet in 1975, we were 80 percent certain that Cygnus was a blackhole. By now I would say that we are about 95 percent certain, but the bet hasyet to be settled.

There is evidence for black holes in a number of other systems in our galaxy,and for much larger black holes at the centers of other galaxies and quasars.One can also consider the possibility that there might be black holes withmasses much less than that of the sun. Such black holes could not be formedby gravitational collapse, because their masses are below the Chandrasekharmass limit. Stars of this low mass can support themselves against the force ofgravity even when they have exhausted their nuclear fuel. So, low-mass blackholes could form only if matter were compressed to enormous densities by verylarge external pressures. Such conditions could occur in a very big hydrogenbomb. The physicist John Wheeler once calculated that if one took all theheavy water in all the oceans of the world, one could build a hydrogen bombthat would compress matter at the center so much that a black hole would becreated. Unfortunately, however, there would be no one left to observe it.A more practical possibility is that such low-mass black holes might have beenformed in the high temperatures and pressures of the very early universe. Blackholes could have been formed if the early universe had not been perfectlysmooth and uniform, because then a small region that was denser than aver-age could be compressed in this way to form a black hole. But we know thatthere must have been some irregularities, because otherwise the matter in theuniverse would still be perfectly uniformly distributed at the present epoch,instead of being clumped together in stars and

galaxies.

Whether or not the irregularities required to account for stars and galaxieswould have led to the formation of a significant number of these primordialblack holes depends on the details of the conditions in the early universe. Soif we could determine how many primordial black holes there are now, wewould learn a lot about the very early stages of the universe. Primordial blackholes with masses more than a thousand million tons-the mass of a largemountain-could be detected only by their gravitational influence on othervisible matter or on the expansion of the universe. However, as we shalllearn in the next lecture, black holes are not really black after all: They glowlike a hot body, and the smaller they are, the more they glow. So, paradoxi-cally, smaller black holes might actually turn out to be easier to detect thanlarge ones.

The Theory of Everything: The Origin and Fate of the Universe

Chapter 4 - FOURTH LECTURE - BLACK HOLES AIN'T SO BLACK

Before 1970, my research on general relativity had concentrated mainly onthe question of whether there had been a big bang singularity. However,one evening in November of that year, shortly after the birth of my daughter,Lucy, I started to think about black holes as I was getting into bed. My disabil-ity made this rather a slow process, so I had plenty of time. At that date therewas no precise definition of which points in space-time lay inside a black holeand which lay outside.

I had already discussed with Roger Penrose the idea of defining a black hole asthe set of events from which it was not possible to escape to a large distance.This is now the generally accepted definition. It means that the boundary ofthe black hole, the event horizon, is formed by rays of light that just fail to getaway from the black hole. Instead, they stay forever, hovering on the edge ofthe black hole. It is like running away from the police and managing to keepone step ahead but not being able to get clear away.

Suddenly I realized that the paths of these light rays could not be approachingone another, because if they were, they must eventually run into each other. Itwould be like someone else running away from the police in the opposite direc-tion. You would both be caught or, in this case, fall into a black hole. But ifthese light rays were swallowed up by the black hole, then they could not havebeen on the boundary of the black hole. So light rays in the event horizon hadto be moving parallel to, or away from, each other.

Another way of seeing this is that the event horizon, the boundary of the blackhole, is like the edge of a shadow. It is the edge of the light of escape to a greatdistance, but, equally, it is the edge of the shadow of impending doom. And ifyou look at the shadow cast by a source at a great distance, such as the sun, youwill see that the rays of light on the edge are not approaching each other. Ifthe rays of light that form the event horizon, the boundary of the black hole,can never approach each other, the area of the event horizon could stay thesame or increase with time. It could never decrease, because that would meanthat at least some of the rays of light in the boundary would have to beapproaching each other. In fact, the area would increase whenever matter orradiation fell into the black hole.

Also, suppose two black holes collided and merged together to form a singleblack hole. Then the area of the event horizon of the final black hole wouldbe greater than the sum of the areas of the event horizons of the original blackholes. This nondecreasing property of the event horizon's area placed animportant restriction on the possible behavior of black holes. I was so excitedwith my discovery that I did not get much sleep that night.The next day I rang up Roger Penrose. He agreed with me. I think, in fact, thathe had been aware of this property of the area. However, he had been using aslightly different definition of a black hole. He had not realized that theboundaries of the black hole according to the two definitions would be thesame, provided the black hole had settled down to a stationary state.

THE SECOND LAW OFTHERMODYNAMICS

The nondecreasing behavior of a black hole's area was very reminiscent of thebehavior of a physical quantity called entropy, which measures the degree ofdisorder of a system. It is a matter of common experience that disorder willtend to increase if things are left to themselves; one has only to leave a housewithout repairs to see that. One can create order out of disorder-for example,one can paint the house. However, that requires expenditure of energy, and sodecreases the amount of ordered energy available.

A precise statement of this idea is known as the second law of thermodynam-ics. It states that the entropy of an isolated system never decreases with time.Moreover, when two systems are joined together, the entropy of the combinedsystem is greater than the sum of the entropies of the individual systems. Forexample, consider a system of gas molecules in a box. The molecules can bethought of as little billiard balls continually colliding with each other andbouncing off the walls of the box. Suppose that initially the molecules are allconfined to the left-hand side of the box by a partition. If the partition is thenremoved, the molecules will tend to spread out and occupy both halves of thebox. At some later time they could, by chance, all be in the right half or all beback in the left half. However, it is overwhelmingly more probable that therewill be roughly equal numbers in the two halves. Such a state is less ordered,or more disordered, than the original state in which all the molecules were inone half. One therefore says that the entropy of the gas has gone up.

Similarly, suppose one starts with two boxes, one containing oxygen moleculesand the other containing nitrogen molecules. If one joins the boxes togetherand removes the intervening wall, the oxygen and the nitrogen molecules willstart to mix. At a later

time, the most probable state would be to have athoroughly uniform mixture of oxygen and nitrogen molecules throughout thetwo boxes. This state would be less ordered, and hence have more entropy,than the initial state of two separate boxes.

The second law of thermodynamics has a rather different status than that ofother laws of science. Other laws, such as Newton's law of gravity, forexample, are absolute law-that is, they always hold. On the other hand, thesecond law is a statistical law-that is, it does not hold always, just in the vastmajority of cases. The probability of all the gas molecules in our box beingfound in one half of the box at a later time is many millions of millions to one,but it could happen.

However, if one has a black hole around, there seems to be a rather easier wayof violating the second law: Just throw some matter with a lot of entropy, suchas a box of gas, down the black hole. The total entropy of matter outside theblack hole would go down. One could, of course, still say that the total entropy,including the entropy inside the black hole, has not gone down. But sincethere is no way to look inside the black hole, we cannot see how much entropythe matter inside it has. It would be nice, therefore, if there was some featureof the black hole by which observers outside the black hole could tell itsentropy; this should increase whenever matter carrying entropy fell into theblack hole.

Following my discovery that the area of the event horizon increased whenevermatter fell into a black hole, a research student at Princeton named JacobBekenstein suggested that the area of the event horizon was a measure of theentropy of the black hole. As matter carrying entropy fell into the black hole,the area of the event horizon would go up, so that the sum of the entropy ofmatter outside black holes and the area of the horizons would never go down.This suggestion seemed to prevent the second law of thermodynamics frombeing violated in most situations. However, there was one fatal flaw: If a blackhole has entropy, then it ought also to have a temperature. But a body with anonzero temperature must emit radiation at a certain rate. It is a matter ofcommon experience that if one heats up a poker in the fire, it glows red hotand emits radiation. However, bodies at lower temperatures emit radiation,too; one just does not normally notice it because the amount is fairly small.This radiation is required in order to prevent violations of the second law. Soblack holes ought to emit radiation, but by their very definition, black holesare objects that are not supposed to emit anything. It therefore seemed that thearea of the event horizon of a black hole

could not be regarded as its entropy.In fact, in 1972 I wrote a paper on this subject with Brandon Carter and anAmerican colleague, Jim Bardeen. We pointed out that, although there weremany similarities between entropy and the area of the event horizon, there wasthis apparently fatal difficulty. I must admit that in writing this paper I wasmotivated partly by irritation with Bekenstein, because I felt he had misusedmy discovery of the increase of the area of the event horizon. However, itturned out in the end that he was basically correct, though in a manner he hadcertainly not expected.

BLACK HOLE RADIATION

In September 1973, while I was visiting Moscow, I discussed black holes withtwo leading Soviet experts, Yakov Zeldovich and Alexander Starobinsky. Theyconvinced me that, according to the quantum mechanical uncertainty princi-ple, rotating black holes should create and emit particles. I believed their argu-ments on physical grounds, but I did not like the mathematical way in whichthey calculated the emission. I therefore set about devising a better mathemat-ical treatment, which I described at an informal seminar in Oxford at the endof November 1973. At that time I had not done the calculations to find outhow much would actually be emitted. I was expecting to discover just the radi-ation that Zeldovich and Starobinsky had predicted from rotating black holes.However, when I did the calculation, I found, to my surprise and annoyance,that even nonrotating black holes should apparently create and emit particlesat a steady rate.

At first I thought that this emission indicated that one of the approximationsI had used was not valid. I was afraid if Bekenstein found out about it, he woulduse it as a further argument to support his ideas about the entropy of blackholes, which I still did not like. However, the more I thought about it, themore it seemed that the approximations really ought to hold. But what finallyconvinced me that the emission was real was that the spectrum of the emittedparticles was exactly that which would be emitted by a hot body.

The black hole was emitting particles at exactly the correct rate to preventviolations of the second law.

Since then, the calculations have been repeated in a number of different formsby other people. They all confirm that a black hole ought to emit particles andradiation as if it were a hot body with a temperature that depends only on theblack hole's mass: the higher the mass, the lower the temperature. One canunderstand this emission in the following way: What we think of as emptyspace cannot be completely empty because that would

mean that all the fields,such as the gravitational field and the electromagnetic field, would have to beexactly zero. However, the value of a field and its rate of change with time arelike the position and velocity of a particle. The uncertainty principle impliesthat the more accurately one knows one of these quantities, the less accuratelyone can know the other.

So in empty space the field cannot be fixed at exactly zero, because then itwould have both a precise value, zero, and a precise rate of change, also zero.Instead, there must be a certain minimum amount of uncertainty, or quantumfluctuations, in the value of a field. One can think of these fluctuations as pairsof particles of light or gravity that appear together at some time, move apart,and then come together again and annihilate each other. These particles arecalled virtual particles. Unlike real particles, they cannot be observed directly

with a particle detector. However, their indirect effects, such as small changesin the energy of electron orbits and atoms, can be measured and agree with thetheoretical predictions to a remarkable degree of accuracy.

By conservation of energy, one of the partners in a virtual particle pair willhave positive energy and the other partner will have negative energy. The onewith negative energy is condemned to be a short-lived virtual particle. This isbecause real particles always have positive energy in normal situations. It musttherefore seek out its partner and annihilate it. However, the gravitationalfield inside a black hole is so strong that even a real particle can have negativeenergy there.

It is therefore possible, if a black hole is present, for the virtual particle withnegative energy to fall into the black hole and become a real particle. In thiscase it no longer has to annihilate its partner; its forsaken partner may fall intothe black hole as well. But because it has positive energy, it is also possible forit to escape to infinity as a real particle. To an observer at a distance, it willappear to have been emitted from the black hole. The smaller the black hole,the less far the particle with negative energy will have to go before it becomesa real particle. Thus, the rate of emission will be greater, and the apparent tem-perature of the black hole will be higher.

The positive energy of the outgoing radiation would be balanced by a flow ofnegative energy particles into the black hole. By Einstein's famous equationE = mc2, energy is equivalent to mass. A flow of negative energy into the blackhole therefore reduces its mass. As the black hole loses mass, the area of itsevent horizon gets smaller, but this decrease in the entropy of the black

holeis more than compensated for by the entropy of the emitted radiation, so thesecond law is never violated.

BLACK HOLE EXPLOSIONS

The lower the mass of the black hole, the higher its temperature is. So as theblack hole loses mass, its temperature and rate of emission increase. It there-fore loses mass more quickly. What happens when the mass of the black holeeventually becomes extremely small is not quite clear. The most reasonableguess is that it would disappear completely in a tremendous final burst of emis-sion, equivalent to the explosion of millions of H-bombs.

A black hole with a mass a few times that of the sun would have a tempera-ture of only one ten-millionth of a degree above absolute zero. This is muchless than the temperature of the microwave radiation that fills the universe,about 2.7 degrees above absolute zero-so such black holes would give off lessthan they absorb, though even that would be very little. If the universe is des-

tined to go on expanding forever, the temperature of the microwave radiationwill eventually decrease to less than that of such a black hole. The hole willthen absorb less than it emits and will begin to lose mass. But, even then, itstemperature is so low that it would take about 1066years to evaporatecompletely. This is much longer than the age of the universe, which is onlyabout 1010 years.

On the other hand, as we learned in the last lecture, there might be primor-dial black holes with a very much smaller mass that were made by the collapseof irregularities in the very early stages of the universe. Such black holes wouldhave a much higher temperature and would be emitting radiation at a muchgreater rate. A primordial black hole with an initial mass of a thousand mil-lion tons would have a lifetime roughly equal to the age of the universe.Primordial black holes with initial masses less than this figure would alreadyhave completely evaporated. However, those with slightly greater masseswould still be emitting radiation in the form of X rays and gamma rays. Theseare like waves of light, but with a much shorter wavelength. Such holeshardly deserve the epithet black. They really are white hot, and are emittingenergy at the rate of about ten thousand megawatts.

One such black hole could run ten large power stations, if only we could har-ness its output. This would be rather difficult, however. The black hole wouldhave the mass of a mountain compressed into the size of the nucleus of anatom. If you had one of these black holes on the surface of the Earth, therewould be no

way to stop it falling through the floor to the center of the Earth.It would oscillate through the Earth and back, until eventually it settled downat the center. So the only place to put such a black hole, in which one mightuse the energy that it emitted, would be in orbit around the Earth. And theonly way that one could get it to orbit the Earth would be to attract it thereby towing a large mass in front of it, rather like a carrot in front of a donkey.This does not sound like a very practical proposition, at least not in theimmediate future.

THE SEARCH FOR PRIMORDIALBLACK HOLES

But even if we cannot harness the emission from these primordial black holes,what are our chances of observing them? We could look for the gamma raysthat the primordial black holes emit during most of their lifetime. Althoughthe radiation from most would be very weak because they are far away, thetotal from all of them might be detectable. We do, indeed, observe such abackground of gamma rays. However, this background was probably generatedby processes other than primordial black holes. One can say that the observa-tions of the gamma ray background do not provide any positive evidence forprimordial black holes. But they tell us that, on average, there cannot be morethan three hundred little black holes in every cubic light-year in the universe.This limit means that primordial black holes could make up at most one mil-lionth of the average mass density in the universe.

With primordial black holes being so scarce, it might seem unlikely that therewould be one that was near enough for us to observe on its own. But sincegravity would draw primordial black holes toward any matter, they should bemuch more common in galaxies. If they were, say, a million times more com-mon in galaxies, then the nearest black hole to us would probably be at adistance of about a thousand million kilometers, or about as far as Pluto, thefarthest known planet. At this distance it would still be very difficult to detectthe steady emission of a black hole even if it was ten thousand megawatts.In order to observe a primordial black hole, one would have to detect severalgamma ray quanta coming from the same direction within a reasonable spaceof time, such as a week.

Otherwise, they might simply be part of the background. But Planck's quan-tum principle tells us that each gamma ray quantum has a very high energy,because gamma rays have a very high frequency. So to radiate even ten thou-sand megawatts would not take many quanta. And to observe these few quan-ta coming from the distance of Pluto would require a larger gamma

ray detec-tor than any that have been constructed so far. Moreover, the detector wouldhave to be in space, because gamma rays cannot penetrate the atmosphere.

Of course, if a black hole as close as Pluto were to reach the end of its life andblow up, it would be easy to detect the final burst of emission. But if the blackhole has been emitting for the last ten or twenty thousand million years, thechances of it reaching the end of its life within the next few years are reallyrather small. It might equally well be a few million years in the past or future.So in order to have a reasonable chance of seeing an explosion before yourresearch grant ran out, you would have to find a way to detect any explosionswithin a distance of about one light-year. You would still have the problem ofneeding a large gamma ray detector to observe several gamma ray quanta fromthe explosion. However, in this case, it would not be necessary to determinethat all the quanta came from the same direction. It would be enough toobserve that they all arrived within a very short time interval to be reasonablyconfident that they were coming from the same burst.

One gamma ray detector that might be capable of spotting primordial blackholes is the entire Earth's atmosphere. (We are, in any case, unlikely to be ableto build a larger detector.) When a high-energy gamma ray quantum hits theatoms in our atmosphere, it creates pairs of electrons and positrons. Whenthese hit other atoms, they in turn create more pairs of electrons and positrons.So one gets what is called an electron shower. The result is a form of lightcalled Cerenkov radiation. One can therefore detect gamma ray bursts bylooking for flashes of light in the night sky.

Of course, there are a number of other phenomena, such as lightning, whichcan also give flashes in the sky. However, one could distinguish gamma raybursts from such effects by observing flashes simultaneously at two or morethoroughly widely separated locations. A search like this has been carried outby two scientists from Dublin, Neil Porter and Trevor Weekes, using telescopesin Arizona. They found a number of flashes but none that could be definitelyascribed to gamma ray bursts from primordial black holes.

Even if the search for primordial black holes proves negative, as it seems itmay, it will still give us important information about the very early stages ofthe universe. If the early universe had been chaotic or irregular, or if the pres-sure of matter had been low, one would have expected it to produce manymore primordial black holes than the limit set by our observations of

thegamma ray background. It is only if the early universe was very smooth anduniform, and with a high pressure, that one can explain the absence ofobservable numbers of primordial black holes.

GENERAL RELATIVITY ANDQUANTUM MECHANICS

Radiation from black holes was the first example of a prediction that depend-ed on both of the great theories of this century, general relativity and quantummechanics. It aroused a lot of opposition initially because it upset the existingviewpoint: “How can a black hole emit anything?” When I first announced theresults of my calculations at a conference at the Rutherford Laboratory nearOxford, I was greeted with general incredulity. At the end of my talk the chair-man of the session, John G. Taylor from Kings College, London, claimed it wasall nonsense. He even wrote a paper to that effect.

However, in the end most people, including John Taylor, have come to theconclusion that black holes must radiate like hot bodies if our other ideasabout general relativity and quantum mechanics are correct. Thus eventhough we have not yet managed to find a primordial black hole, there isfairly general agreement that if we did, it would have to be emitting a lot ofgamma and X rays. If we do find one, I will get the Nobel Prize.The existence of radiation from black holes seems to imply that gravitationalcollapse is not as final and irreversible as we once thought. If an astronaut fallsthat extra mass will be returned to the universe in the form of radiation. Thus,in a sense, the astronaut will be recycled. It would be a poor sort of immortal-ity, however, because any personal concept of time for the astronaut wouldalmost certainly come to an end as he was crushed out of existence inside theblack hole. Even the types of particle that were eventually emitted by theblack hole would in general be different from those that made up the astro-naut. The only feature of the astronaut that would survive would be his massor energy.

The approximations I used to derive the emission from black holes shouldwork well when the black hole has a mass greater than a fraction of a gram.However, they will break down at the end of the black hole’s life, when itsmass gets very small. The most likely outcome seems to be that the black holewould just disappear, at least from our region of the universe. It would takewith it the astronaut and any singularity there might be inside the black hole.This was the first indication that quantum mechanics might remove the sin-gularities that were predicted by classical general relativity. However, themethods that I and other

people were using in 1974 to study the quantumeffects of gravity were not able to answer questions such as whether singulari-ties would occur in quantum gravity.

From 1975 onward, I therefore started to develop a more powerful approach toquantum gravity based on Feynman's idea of a sum over histories. The answersthat this approach suggests for the origin and fate of the universe will bedescribed in the next two lectures. We shall see that quantum mechanicsallows the universe to have a beginning that is not a singularity. This meansthat the laws of physics need not break down at the origin of the universe. Thestate of the universe and its contents, like ourselves, are completely deter-mined by the laws of physics, up to the limit set by the uncertainty principle.So much for free will.

The Theory of Everything: The Origin and Fate of the Universe

Chapter 5 - FIFTH LECTURE - THE ORIGIN AND FATE OF THE...

T H E O R I G I N A N D F A T E O F T H E U N I V E R S

EThroughout the 1970s I had been working mainly on black holes. However,n 1981 my interest in questions about the origin of the universe wasreawakened when I attended a conference on cosmology in the Vatican. TheCatholic church had made a bad mistake with Galileo when it tried to laydown the law on a question of science, declaring that the sun went around theEarth. Now, centuries later, it had decided it would be better to invite a num-ber of experts to advise it on cosmology.

At the end of the conference the participants were granted an audience withthe pope. He told us that it was okay to study the evolution of the universeafter the big bang, but we should not inquire into the big bang itself becausethat was the moment of creation and therefore the work of God.I was glad then that he did not know the subject of the talk I had just given atthe conference. I had no desire to share the fate of Galileo; I have a lot of sym-pathy with Galileo, partly because I was born exactly three hundred years afterhis death.

THE HOT BIG BANG MODEL

In order to explain what my paper was about, I shall first describe the generallyaccepted history of the universe, according to what is known as the “hot bigbang model.” This assumes that the universe is described by a Friedmannmodel, right back to the big bang. In such models one finds that as the uni-verse expands, the temperature of the matter and radiation in it will go down.Since temperature is simply a measure of the average energy of the particles,this cooling of the universe will have a major effect on the matter in it. At veryhigh temperatures, particles will be moving around so fast that they can escapeany attraction toward each other caused by the nuclear or electromagneticforces. But as they cooled off, one would expect particles that attract eachother to start to clump together.

At the big bang itself, the universe had zero size and so must have been infi-nitely hot. But as the universe expanded, the temperature of the radiationwould have decreased. One second after the big bang it would have fallen toabout ten thousand million degrees. This is about a thousand times the tem-perature at the center of the sun, but temperatures as high as this are reachedin H-bomb explosions. At this time the universe would

have contained mostlyphotons, electrons, and neutrinos and their antiparticles, together with someprotons and neutrons.

As the universe continued to expand and the temperature to drop, the rate atwhich electrons and the electron pairs were being produced in collisions wouldhave fallen below the rate at which they were being destroyed by annihilation.So most of the electrons and antielectrons would have annihilated each otherto produce more photons, leaving behind only a few electrons.

About one hundred seconds after the big bang, the temperature would havefallen to one thousand million degrees, the temperature inside the hotteststars. At this temperature, protons and neutrons would no longer have suffi-cient energy to escape the attraction of the strong nuclear force. They wouldstart to combine together to produce the nuclei of atoms of deuterium, orheavy hydrogen, which contain one proton and one neutron. The deuteriumnuclei would then have combined with more protons and neutrons to makehelium nuclei, which contained two protons and two neutrons. There wouldalso be small amounts of a couple of heavier elements, lithium and beryllium.One can calculate that in the hot big bang model about a quarter of the pro-tons and neutrons would have been converted into helium nuclei, along witha small amount of heavy hydrogen and other elements. The remaining neu-trons would have decayed into protons, which are the nuclei of ordinaryhydrogen atoms. These predictions agree very well with what is observed.The hot big bang model also predicts that we should be able to observe theradiation left over from the hot early stages. However, the temperature wouldhave been reduced to a few degrees above absolute zero by the expansion of theuniverse. This is the explanation of the microwave background of radiationthat was discovered by Penzias and Wilson in 1965. We are thereforethoroughly confident that we have the right picture, at least back to about onesecond after the big bang. Within only a few hours of the big bang, theproduction of helium and other elements would have stopped. And after that,for the next million years or so, the universe would have just continuedexpanding, without anything much happening. Eventually, once the tempera-ture had dropped to a few thousand degrees, the electrons and nuclei would nolonger have had enough energy to overcome the electromagnetic attractionbetween them. They would then have started combining to form atoms.

The universe as a whole would have continued expanding and cooling.However, in regions that were slightly denser than average, the expansionwould have been slowed down by extra

gravitational attraction. This wouldeventually stop expansion in some regions and cause them to start to recol-lapse. As they were collapsing, the gravitational pull of matter outside theseregions might start them rotating slightly. As the collapsing region gotsmaller, it would spin faster-just as skaters spinning on ice spin faster as thedraw in their arms. Eventually, when the region got small enough, it would bespinning fast enough to balance the attraction of gravity. In this way, disklikerotating galaxies were born.

As time went on, the gas in the galaxies would break up into smaller cloudsthat would collapse under their own gravity. As these contracted, the temper-ature of the gas would increase until it became hot enough to start nuclearreactions. These would convert the hydrogen into more helium, and the heatgiven off would raise the pressure, and so stop the clouds from contracting anyfurther. They would remain in this state for a long time as stars like our sun,burning hydrogen into helium and radiating the energy as heat and light.More massive stars would need to be hotter to balance their stronger gravita-tional attraction. This would make the nuclear fusion reactions proceed somuch more rapidly that they would use up their hydrogen in as little as a hun-dred million years. They would then contract slightly and, as they heated upfurther, would start to convert helium into heavier elements like carbon oroxygen. This, however, would not release much more energy, so a crisis wouldoccur, as I described in my lecture on black holes.

What happens next is not completely clear, but it seems likely that the centralregions of the star would collapse to a very dense state, such as a neutron staror black hole. The outer regions of the star may get blown off in a tremendousexplosion called a supernova, which would outshine all the other stars in thegalaxy. Some of the heavier elements produced near the end of the star's lifewould be flung back into the gas in the galaxy. They would provide some ofthe raw material for the next generation of stars.

Our own sun contains about 2 percent of these heavier elements because it isa second- or third-generation star. It was formed some five thousand millionyears ago out of a cloud of rotating gas containing the debris of earlier super-novas. Most of the gas in that cloud went to form the sun or got blown away.However, a small amount of the heavier elements collected together to formthe bodies that now orbit the sun as planets like the Earth.

OPEN QUESTIONS

This picture of a universe that started off very hot and cooled

as it expanded isin agreement with all the observational evidence that we have today.Nevertheless, it leaves a number of important questions unanswered. First, whywas the early universe so hot? Second, why is the universe so uniform on a largescale-why does it look the same at all points of space and in all directions?Third, why did the universe start out with so nearly the critical rate of expan-sion to just avoid recollapse? If the rate of expansion one second after the bigbang had been smaller by even one part in a hundred thousand millionmillion, the universe would have recollapsed before it ever reached its presentsize. On the other hand, if the expansion rate at one second had been largerby the same amount, the universe would have expanded so much that it wouldbe effectively empty now.

Fourth, despite the fact that the universe is so uniform and homogenous on alarge scale, it contains local lumps such as stars and galaxies. These are thoughtto have developed from small differences in the density of the early universefrom one region to another. What was the origin of these density fluctuations?The general theory of relativity, on its own, cannot explain these features oranswer these questions. This is because it predicts that the universe started offwith infinite density at the big bang singularity. At the singularity, general rel-ativity and all other physical laws would break down. One cannot predict whatwould come out of the singularity. As I explained before, this means that onemight as well cut any events before the big bang out of the theory, because theycan have no effect on what we observe. Space-time would have a boundary-a beginning at the big bang. Why should the universe have started off at thebig bang in just such a way as to lead to the state we observe today? Why is theuniverse so uniform, and expanding at just the critical rate to avoid recollapse?One would feel happier about this if one could show that quite a number ofdifferent initial configurations for the universe would have evolved to producea universe like the one we observe.

If this is the case, a universe that developed from some sort of random initialconditions should contain a number of regions that are like what we observe.There might also be regions that were very different. However, these regionswould probably not be suitable for the formation of galaxies and stars. Theseare essential prerequisites for the development of intelligent life, at least as weknow it. Thus, these regions would not contain any beings to observe that theywere different.

When one considers cosmology, one has to take into account the selectionprinciple that we live in a region of the universe that

is suitable for intelligentlife. This fairly obvious and elementary consideration is sometimes called theanthropic principle. Suppose, on the other hand, that the initial state of theuniverse had to be chosen extremely carefully to lead to something like whatwe see around us. Then the universe would be unlikely to contain any regionin which life would appear.

In the hot big bang model that I described earlier, there was not enough timein the early universe for heat to have flowed from one region to another. Thismeans that different regions of the universe would have had to have startedout with exactly the same temperature in order to account for the fact that themicrowave background has the same temperature in every direction we look.Also, the initial rate of expansion would have had to be chosen very preciselyfor the universe not to have recollapsed before now. This means that the ini-tial state of the universe must have been very carefully chosen indeed if thehot big bang model was correct right back to the beginning of time. It wouldbe very difficult to explain why the universe should have begun in just thisway, except as the act of a God who intended to create beings like us.

THE INFLATIONARY MODEL

In order to avoid this difficulty with the very early stages of the hot big bangmodel, Alan Guth at the Massachusetts Institute of Technology put forward anew model. In this, many different initial configurations could have evolved tosomething like the present universe. He suggested that the early universe mighthave had a period of very rapid, or exponential, expansion. This expansion issaid to be inflationary-an analogy with the inflation in prices that occurs to agreater or lesser degree in every country. The world record for price inflationwas probably in Germany after the first war, when the price of a loaf of breadwent from under a mark to millions of marks in a few months. But the inflationwe think may have occurred in the size of the universe was much greater eventhan that-a million million million million million million times in only a tiny frac-tion of a second. Of course, that was before the present government.

Guth suggested that the universe started out from the big bang very hot. Onewould expect that at such high temperatures, the strong and weak nuclearforces and the electromagnetic force would all be unified into a single force.As the universe expanded, it would cool, and particle energies would go down.Eventually there would be what is called a phase transition, and the symmetrybetween the forces would be broken. The strong force would become differentfrom the weak and electromagnetic forces.

One common example of a phasetransition is the freezing of water when you cool it down. Liquid water is sym-metrical, the same at every point and in every direction. However, when icecrystals form, they will have definite positions and will be lined up in somedirection. This breaks the symmetry of the water.

In the case of water, if one is careful, one can “supercool” it. That is, one canreduce the temperature below the freezing point-0 degrees centigrade-with-out ice forming. Guth suggested that the universe might behave in a similarway: The temperature might drop below the critical value without the symme-try between the forces being broken. If this happened, the universe would bein an unstable state, with more energy than if the symmetry had been broken.This special extra energy can be shown to have an antigravitational effect. Itwould act just like a cosmological constant.

Einstein introduced the cosmological constant into general relativity when hewas trying to construct a static model of the universe. However,in this case,the universe would already be expanding. The repulsive effect of this cosmo-logical constant would therefore have made the universe expand at an ever-increasing rate. Even in regions where there were more matter particles thanaverage, the gravitational attraction of the matter would have been out-weighed by the repulsion of the effective cosmological constant. Thus, theseregions would also expand in an accelerating inflationary manner.

As the universe expanded, the matter particles got farther apart. One would beleft with an expanding universe that contained hardly any particles. It wouldstill be in the supercooled state, in which the symmetry between the forces isnot broken. Any irregularities in the universe would simply have beensmoothed out by the expansion, as the wrinkles in a balloon are smoothedaway when you blow it up. Thus, the present smooth and uniform state of theuniverse could have evolved from many different nonuniform initial states.The rate of expansion would also tend toward just the critical rate needed toavoid recollapse.

Moreover, the idea of inflation could also explain why there is so much matterin the universe. There are something like 1,080 particles in the region of theuniverse that we can observe. Where did they all come from? The answer isthat, in quantum theory, particles can be created out of energy in the form ofparticle/antiparticle pairs. But that just raises the question of where the energycame from. The answer is that the total energy of the universe is exactly zero.The matter in the universe is made out of

positive energy. However, the mat-ter is all attracting itself by gravity. Two pieces of matter that are close to eachother have less energy than the same two pieces a long way apart. This isbecause you have to expend energy to separate them. You have to pull againstthe gravitational force attracting them together. Thus, in a sense, the gravita-tional field has negative energy. In the case of the whole universe, one canshow that this negative gravitational energy exactly cancels the positive ener-gy of the matter. So the total energy of the universe is zero.

Now, twice zero is also zero. Thus, the universe can double the amount of pos-itive matter energy and also double the negative gravitational energy withoutviolation of the conservation of energy. This does not happen in the normalexpansion of the universe in which the matter energy density goes down as theuniverse gets bigger. It does happen, however, in the inflationary expansion,because the energy density of the supercooled state remains constant while theuniverse expands. When the universe doubles in size, the positive matter ener-gy and the negative gravitational energy both double, so the total energyvery large amount. Thus, the total amount of energy available to make parti-cles becomes very large. As Guth has remarked, “It is said that there is no suchthing as a free lunch. But the universe is the ultimate free lunch.”

THE END OF INFLATION

The universe is not expanding in an inflationary way today. Thus, there hadto be some mechanism that would eliminate the very large effective cosmolog-ical constant. This would change the rate of expansion from an acceleratedone to one that is slowed down by gravity, as we have today. As the universeexpanded and cooled, one might expect that eventually the symmetry betweenthe forces would be broken, just as supercooled water always freezes in the end.The extra energy of the unbroken symmetry state would then be released andwould reheat the universe. The universe would then go on to expand and cool,just like the hot big bang model. However, there would now be an explanationof why the universe was expanding at exactly the critical rate and why differ-ent regions had the same temperature.

In Guth’s original proposal, the transition to broken symmetry was supposed tooccur suddenly, rather like the appearance of ice crystals in very cold water.The idea was that “bubbles” of the new phase of broken symmetry would haveformed in the old phase, like bubbles of steam surrounded by boiling water.The bubbles were supposed to expand and meet up with each other until

thewhole universe was in the new phase. The trouble was, as I and several otherpeople pointed out, the universe was expanding so fast that the bubbles wouldbe moving away from each other too rapidly to join up. The universe would beleft in a very nonuniform state, with some regions having symmetry betweenthe different forces. Such a model of the universe would not correspond towhat we see.

In October 1981 I went to Moscow for a conference on quantum gravity. Afterthe conference, I gave a seminar on the inflationary model and its problems atthe Sternberg Astronomical Institute. In the audience was a young Russian,Andrei Linde. He said that the difficulty with the bubbles not joining up couldbe avoided if the bubbles were very big. In this case, our region of the universecould be contained inside a single bubble. In order for this to work, the changefrom symmetry to broken symmetry must have taken place very slowly insidethe bubble, but this is quite possible according to grand unified theories.Linde's idea of a slow breaking of symmetry was very good, but I pointed outthat his bubbles would have been bigger than the size of the universe at thetime. I showed that instead the symmetry would have broken everywhere atthe same time, rather than just inside bubbles. This would lead to a uniformuniverse, like we observe. The slow symmetry breaking model was a goodattempt to explain why the universe is the way it is. However, I and severalother people showed that it predicted much greater variations in themicrowave background radiation than are observed. Also, later work castdoubt on whether there would have been the right kind of phase transition inthe early universe. A better model, called the chaotic inflationary model, wasintroduced by Linde in 1983. This doesn't depend on phase transitions, and itcan give us the right size of variations of the microwave background. The infla-tionary model showed that the present state of the universe could have arisenfrom quite a large number of different initial configurations. It cannot be thecase, however, that every initial configuration would have led to a universelike the one we observe. So even the inflationary model does not tell us whythe initial configuration was such as to produce what we observe. Must we turnto the anthropic principle for an explanation? Was it all just a lucky chance?That would seem a counsel of despair, a negation of all our hopes of under-standing the underlying order of the universe.

QUANTUM GRAVITY

In order to predict how the universe should have started off, one needs laws thathold at the beginning of time. If the classical theory of general relativity wascorrect, the singularity theorem

showed that the beginning of time would havebeen a point of infinite density and curvature. All the known laws of sciencewould break down at such a point. One might suppose that there were new lawsthat held at singularities, but it would be very difficult even to formulate lawsat such badly behaved points and we would have no guide from observations asto what those laws might be. However, what the singularity theorems reallyindicate is that the gravitational field becomes so strong that quantum gravita-tional effects become important: Classical theory is no longer a good descrip-tion of the universe. So one has to use a quantum theory of gravity to discussthe very early stages of the universe. As we shall see, it is possible in the quan-tum theory for the ordinary laws of science to hold everywhere, including at thebeginning of time. It is not necessary to postulate new laws for singularities,because there need not be any singularities in the quantum theory.

We don't yet have a complete and consistent theory that combines quantummechanics and gravity. However, we are thoroughly certain of some featuresthat such a unified theory should have. One is that it should incorporateFeynman's proposal to formulate quantum theory in terms of a sum over histo-ries. In this approach, a particle going from A to B does not have just a singlehistory as it would in a classical theory. Instead, it is supposed to follow everypossible path in space-time. With each of these histories, there are associateda couple of numbers, one representing the size of a wave and the other repre-senting its position in the cycle-its phase.The probability that the particle, say, passes through some particular point isfound by adding up the waves associated with every possible history thatpasses through that point. When one actually tries to perform these sums,however, one runs into severe technical problems. The only way around theseis the following peculiar prescription: One must add up the waves for particlehistories that are not in the real time that you and I experience but take placein imaginary time.

Imaginary time may sound like science fiction, but it is in fact a well-definedmathematical concept. To avoid the technical difficulties with Feynman's sumover histories, one must use imaginary time. This has an interesting effect onspace-time: The distinction between time and space disappears completely. Aspace-time in which events have imaginary values of the time coordinate issaid to be Euclidean because the metric is positive definite.In Euclidean space-time there is no difference between the time direction anddirections in space. On the other hand, in real space-time, in which events arelabeled by real values of the

time coordinate, it is easy to tell the difference. Thetime direction lies within the light cone, and space directions lie outside. Onecan regard the use of imaginary time as merely a mathematical device-ortrick-to calculate answers about real space-time. However, there may be moreto it than that. It may be that Euclidean space-time is the fundamental conceptand what we think of as real space-time is just a figment of our imagination.When we apply Feynman's sum over histories to the universe, the analogue ofthe history of a particle is now a complete curved space-time which representsthe history of the whole universe. For the technical reasons mentioned above,these curved space-times must be taken to be Euclidean. That is, time isimaginary and is indistinguishable from directions in space. To calculate theprobability of finding a real space-time with some certain property, one addsup the waves associated with all the histories in imaginary time that have thatproperty. One can then work out what the probable history of the universewould be in real time.

THE NO BOUNDARY CONDITION

In the classical theory of gravity, which is based on real space-time, there areonly two possible ways the universe can behave. Either it has existed for an infi-nite time, or else it had a beginning at a singularity at some finite time in thepast. In fact, the singularity theorems show it must be the second possibility. Inthe quantum theory of gravity, on the other hand, a third possibility arises.Because one is using Euclidean space-times, in which the time direction is onthe same footing as directions in space, it is possible for space-time to be finitein extent and yet to have no singularities that formed a boundary or edge.Space-time would be like the surface of the Earth, only with two more dimensions. The surface of the Earth is finite in extent but it doesn't have a boundaryor edge. If you sail off into the sunset, you don't fall off the edge or run into asingularity. I know, because I have been around the world.

If Euclidean space-times direct back to infinite imaginary time or else startedat a singularity, we would have the same problem as in the classical theory ofspecifying the initial state of the universe. God may know how the universebegan, but we cannot give any particular reason for thinking it began one wayrather than another. On the other hand, the quantum theory of gravity hasopened up a new possibility. In this, there would be no boundary tospace-time. Thus, there would be no need to specify the behavior at theboundary. There would be no singularities at which the laws of science brokedown and no edge of space-time

at which one would have to appeal to God orsome new law to set the boundary conditions for space-time. One could say:"The boundary condition of the universe is that it has no boundary." The uni-verse would be completely self-contained and not affected by anything outsideitself. It would be neither created nor destroyed. It would just be.

It was at the conference in the Vatican that I first put forward the suggestionthat maybe time and space together formed a surface that was finite in size butdid not have any boundary or edge. My paper was rather mathematical, how-ever, so its implications for the role of God in the creation of the universe werenot noticed at the time-just as well for me. At the time of the Vatican confer-ence, I did not know how to use a no boundary idea to make predictions aboutthe universe. However, I spent the following summer at the University ofCalifornia, Santa Barbara. There, a friend and colleague of mine, Jim Hartle,worked out with me what conditions the universe must satisfy if space-timehad no boundary.

I should emphasize that this idea that time and space should be finite withoutboundary is just a proposal. It cannot be deduced from some other principle.Like any other scientific theory, it may initially be put forward for aesthetic ormetaphysical reasons, but the real test is whether it makes predictions thatagree with observation. This, however, is difficult to determine in the case ofquantum gravity, for two reasons. First, we are not yet sure exactly which the-ory successfully combines general relativity and quantum mechanics, thoughwe know quite a lot about the form such a theory must have. Second, anymodel that described the whole universe in detail would be much too compli-cated mathematically for us to be able to calculate exact predictions. Onetherefore has to make approximations-and even then, the problem ofextracting predictions remains a difficult one.

One finds, under the no boundary proposal, that the chance of the universebeing found to be following most of the possible histories is negligible. Butthere is a particular family of histories that are much more probable than theothers. These histories may be pictured as being like the surface of the Earth,with a distance from the North Pole representing imaginary time; the size of acircle of latitude would represent the spatial size of the universe. The universestarts at the North Pole as a single point. As one moves south, the circle of lat-itude get bigger, corresponding to the universe expanding with imaginary time.The universe would reach a maximum size at the equator and would contractagain to a single point at the South Pole. Even though the universe

wouldhave zero size at the North and South poles, these points would not be singu-larities any more than the North and South poles on the Earth are singular.The laws of science will hold at the beginning of the universe, just as they doat the North and South poles on the Earth.

The history of the universe in real time, however, would look very different. Itwould appear to start at some minimum size, equal to the maximum size of thehistory in imaginary time. The universe would then expand in real time likethe inflationary model. However, one would not now have to assume that theuniverse was created somehow in the right sort of state. The universe wouldexpand to a very large size, but eventually it would collapse again into whatlooks like a singularity in real time. Thus, in a sense, we are still all doomed,even if we keep away from black holes. Only if we could picture the universein terms of imaginary time would there be no singularities.

The singularity theorems of classical general relativity showed that the uni-verse must have a beginning, and that this beginning must be described interms of quantum theory. This in turn led to the idea that the universe couldbe finite in imaginary time, but without boundaries or singularities. When onegoes back to the real time in which we live, however, there will still appear tobe singularities. The poor astronaut who falls into a black hole will still cometo a sticky end. It is only if he could live in imaginary time that he wouldencounter no singularities.

This might suggest that the so-called imaginary time is really the fundamen-tal time, and that what we call real time is something we create just in ourminds. In real time, the universe has a beginning and an end at singularitiesthat form a boundary to space-time and at which the laws of science breakdown. But in imaginary time, there are no singularities or boundaries. Somaybe what we call imaginary time is really more basic, and what we call realtime is just an idea that we invent to help us describe what we think the uni-verse is like. But according to the approach I described in the first lecture, ascientific theory is just a mathematical model we make to describe our obser-vations. It exists only in our minds. So it does not have any meaning to ask:Which is real, “real” or “imaginary” time? It is simply a matter of which is amore useful description.

The no boundary proposal seems to predict that, in real time, the universeshould behave like the inflationary models. A particularly interesting problemis the size of the small departures from uniform density in the early universe.These are thought to have led to the formation first of the galaxies, then ofstars, and

finally of beings like us. The uncertainty principle implies that theearly universe cannot have been completely uniform. Instead, there must havebeen some uncertainties or fluctuations in the positions and velocities of theparticles. Using the no boundary condition, one finds that the universe musthave started off with just the minimum possible nonuniformity allowed by theuncertainty principle.

The universe would have then undergone a period of rapid expansion, like inthe inflationary models. During this period, the initial nonuniformities wouldhave been amplified until they could have been big enough to explain the ori-gin of galaxies. Thus, all the complicated structures that we see in the universemight be explained by the no boundary condition for the universe and theuncertainty principle of quantum mechanics.

The idea that space and time may form a closed surface without boundary alsohas profound implications for the role of God in the affairs of the universe.With the success of scientific theories in describing events, most people havecome to believe that God allows the universe to evolve according to a set oflaws. He does not seem to intervene in the universe to break these laws.However, the laws do not tell us what the universe should have looked likewhen it started. It would still be up to God to wind up the clockwork andchoose how to start it off. So long as the universe had a beginning that was asingularity, one could suppose that it was created by an outside agency. But ifthe universe is really completely self-contained, having no boundary or edge,it would be neither created nor destroyed. It would simply be. What place,then, for a creator?

The Theory of Everything: The Origin and Fate of the Universe

Chapter 6 - SIXTH LECTURE - THE DIRECTION OF TIME

T H E D I R E C T I O N O F T I M EIn his book, The Go Between, L. P. Hartley wrote, “The past is a foreigncountry. They do things differently there-but why is the past so differentfrom the future? Why do we remember the past, but not the future?” In otherwords, why does time go forward? Is this connected with the fact that the uni-verse is expanding?

C, P, T

The laws of physics do not distinguish between the past and the future. Moreprecisely, the laws of physics are unchanged under the combination of opera-tions known as C, P, and T. (C means changing particles for antiparticles. Pmeans taking the mirror image so left and right are swapped for each other.And T means reversing the direction of motion of all particles-in effect, run-ning the motion backward.) The laws of physics that govern the behavior ofmatter under all normal situations are unchanged under the operations C andP on their own. In other words, life would be just the same for the inhabitantsof another planet who were our mirror images and who were made of antimat-ter. If you meet someone from another planet and he holds out his left hand,don’t shake it. He might be made of antimatter. You would both disappear ina tremendous flash of light. If the laws of physics are unchanged by the com-bination of operations C and P, and also by the combination C, P, and T, theymust also be unchanged under the operation T alone. Yet, there is a big differ-ence between the forward and backward directions of time in ordinary life.Imagine a cup of water falling off a table and breaking in pieces on the floor.If you take a film of this, you can easily tell whether it is being run forward orbackward. If you run it backward, you will see the pieces suddenly gather them-selves together off the floor and jump back to form a whole cup on the table.You can tell that the film is being run backward because this kind of behavioris never observed in ordinary life. If it were, the crockery manufacturers wouldgo out of business.

THE ARROWS OF TIME

The explanation that is usually given as to why we don’t see broken cups jump-ing back onto the table is that it is forbidden by the second law of thermody-namics. This says that disorder or entropy always increases with time. In otherwords, it is Murphy’s Law-things get worse. An intact cup on the table is astate of high

order, but a broken cup on the floor is a disordered state. One cantherefore go from the whole cup on the table in the past to the broken cup onthe floor in the future, but not the other way around.

The increase of disorder or entropy with time is one example of what is calledan arrow of time, something that gives a direction to time and distinguishes thepast from the future. There are at least three different arrows of time. First,there is the thermodynamic arrow of time-the direction of time in which dis-order or entropy increases. Second, there is the psychological arrow of time.This is the direction in which we feel time passes-the direction of time inwhich we remember the past, but not the future. Third, there is the cosmolog-ical arrow of time. This is the direction of time in which the universe isexpanding rather than contracting.

I shall argue the the pyschological arrow is determined by the thermodynamicarrow and that these two arrows always point in the same direction. If one makesthe no boundary assumption for the universe, they are related to the cosmolog-ical arrow of time, though they may not point in the same direction. However,I shall argue that it is only when they agree with the cosmological arrow thatthere will be intelligent beings who can ask the question: Why does disorderincrease in the same direction of time as that in which the universe expands?

THE THERMODYNAMIC ARROW

I shall talk first about the thermodynamic arrow of time. The second law ofthermodynamics is based on the fact that there are many more disorderedstates than there are ordered ones. For example, consider the pieces of a jigsawin a box. There is one, and only one, arrangement in which the pieces make acomplete picture. On the other hand, there are a very large number of arrange-ments in which the pieces are disordered and don't make a picture.Suppose a systems starts out in one of the small number of ordered states. Astime goes by, the system will evolve according to the laws of physics and itsstate will change. At a later time, there is a high probability that it will be ina more disordered state, simply because there are so many more disorderedstates. Thus, disorder will tend to increase with time if the system obeys an ini-tial condition of high order.

Suppose the pieces of the jigsaw start off in the ordered arrangement in whichthey form a picture. If you shake the box, the pieces will take up anotherarrangement. This will probably be a disordered arrangement in which thepieces don't form a proper picture, simply because there are so many moredisordered

arrangements. Some groups of pieces may still form parts of thepicture, but the more you shake the box, the more likely it is that these groupswill get broken up. The pieces will take up a completely jumbled state in whichthey don't form any sort of picture. Thus, the disorder of the pieces willprobably increase with time if they obey the initial condition that they start ina state of high order.

Suppose, however, that God decided that the universe should finish up at latetimes in a state of high order but it didn't matter what state it started in. Then,at early times the universe would probably be in a disordered state, and disor-der would decrease with time. You would have broken cups gathering them-selves together and jumping back on the table. However, any human beingswho observing the cups would be living in a universe in which disorderdecreased with time. I shall argue that such beings would have a psychologicalarrow of time that was backward. That is, they would remember thence at latetimes and not remember thence at early times.

THE PSYCHOLOGICAL ARROW

It is rather difficult to talk about human memory because we don't know howthe brain works in detail. We do, however, know all about how computermemories work. I shall therefore discuss the psychological arrow of time forcomputers. I think it is reasonable to assume that the arrow for computers isthe same as that for human. If it were not, one could make a killing on thestock exchange by having a computer that would remember tomorrow's prices.A computer memory is basically some device that can be in either one of twostates. An example would be a superconducting loop of wire. If there is an elec-tric current flowing in the loop, it will continue to flow because there is noresistance. On the other hand, if there is no current, the loop will continuewithout a current. One can label the two states of the memory "one" and "zero."Before an item is recorded in the memory, the memory is in a disordered statewith equal probabilities for one and zero. After the memory interacts with thesystem to be remembered, it will definitely be in one state or the other, accord-ing to the state of the system. Thus, the memory passes from a disordered stateto an ordered one. However, in order to make sure that the memory is in theright state, it is necessary to use a certain amount of energy. This energy is dis-sipated as heat and increases the amount of disorder in the universe. One canshow that this increase of disorder is greater than the increase in the order ofthe memory. Thus, when a computer records an item in memory, the totalamount of disorder

in the universe goes up.

The direction of time in which a computer remembers the past is the same asthat in which disorder increases. This means that our subjective sense of thedirection of time, the psychological arrow of time, is determined by the ther-modynamic arrow of time. This makes the second law of thermodynamicsalmost trivial. Disorder increases with time because we measure time in thedirection in which disorder increases. You can't have a safer bet than that.

THE BOUNDARY CONDITIONS OF THE UNIVERSE

But why should the universe be in a state of high order at one end of time, theend that we call the past? Why was it not in a state of complete disorder at alltimes? After all, this might seem more probable. And why is the direction oftime in which disorder increases the same as that in which the universeexpands? One possible answer is that God simply chose that the universeshould be in a smooth and ordered state at the beginning of the expansionphase. We should not try to understand why or question His reasons becausethe beginning of the universe was the work of God. But the whole history ofthe universe can be said to be the work of God.

It appears that the universe evolves according to well-defined laws. These lawsmay or may not be ordained by God, but it seems that we can discover andunderstand them. Is it, therefore, unreasonable to hope that the same or simi-lar laws may also hold at the beginning of the universe? In the classicaltheory of general relativity, the beginning of the universe has to be a singular-ity of infinite density in space-time curvature. Under such conditions, all theknown laws of physics would break down. Thus, one could not use them topredict how the universe would begin.

The universe could have started out in a very smooth and ordered state. Thiswould have led to well-defined thermodynamic and cosmological arrows oftime, like we observe. But it could equally well have started out in a verylumpy and disordered state. In this case, the universe would already be in astate of complete disorder, so disorder could not increase with time. It wouldeither stay constant, in which case there would be no well-defined thermody-namic arrow of time, or it would decrease, in which case the thermodynamicarrow of time would point in the opposite direction to the cosmological arrow.Neither of these possibilities would agree with what we observe.

As I mentioned, the classical theory of general relativity predicts that theuniverse should begin with a singularity where

the curvature of space-time isinfinite. In fact, this means that classical general relativity predicts its owndownfall. When the curvature of space-time becomes large, quantum gravita-tional effects will become important and the classical theory will cease to be agood description of the universe. One has to use the quantum theory ofgravity to understand how the universe began.

In a quantum theory of gravity, one considers all possible histories of theuniverse. Associated with each history, there are a couple of numbers. Onerepresents the size of a wave and the other the face of the wave, that is,whether the wave is at a crest or a trough. The probability of the universehaving a particular property is given by adding up the waves for all the histo-ries with that property. The histories would be curved spaces that wouldrepresent the evolution of the universe in time. One would still have to sayhow the possible histories of the universe would behave at the boundary ofspace-time in the past. We do not and cannot know the boundary conditionsof the universe in the past. However, one could avoid this difficulty if theboundary condition of the universe is that it has no boundary. In other words,all the possible histories are finite in extent but have no boundaries, edges, orsingularities. They are like the surface of the Earth, but with two more dimen-sions. In that case, the beginning of time would be a regular smooth point ofspace-time. This means that the universe would have begun its expansion ina very smooth and ordered state. It could not have been completely uniformbecause that would violate the uncertainty principle of quantum theory. Therehad to be small fluctuations in the density and velocities of particles. The noboundary condition, however, would imply that these fluctuations were assmall as they could be, consistent with the uncertainty principle.

The universe would have started off with a period of exponential or “inflation-ary” expansion. In this, it would have increased its size by a very large factor.During this expansion, the density fluctuations would have remained small atfirst, but later would have started to grow. Regions in which the density wasslightly higher than average would have had their expansion slowed down bythe gravitational attraction of the extra mass. Eventually, such regions wouldstop expanding, and would collapse to form galaxies, stars, and beings like us.The universe would have started in a smooth and ordered state and wouldbecome lumpy and disordered as time went on. This would explain the exis-tence of the thermodynamic arrow of time. The universe would start in a stateof high order and would become more disordered with time. As I showed ear-lier, the

psychological arrow of time points in the same direction as the ther-modynamic arrow. Our subjective sense of time would therefore be that inwhich the universe is expanding, rather than the opposite direction, in whichit would be contracting.

DOES THE ARROW OF TIME REVERSE?

But what would happen if and when the universe stopped expanding andbegan to contract again? Would the thermodynamic arrow reverse anddisorder begin to decrease with time? This would lead to all sorts ofscience-fiction-like possibilities for people who survived from the expandingto the contracting phase. Would they see broken cups gathering themselvestogether off the floor and jumping back on the table? Would they be able toremember tomorrow's prices and make a fortune on the stock market?It might seem a bit academic to worry about what would happen when the uni-verse collapses again, as it will not start to contract for at least another tenthousand million years. But there is a quicker way to find out what will hap-pen: Jump into a black hole. The collapse of a star to form a black hole is ratherlike the later stages of the collapse of the whole universe. Thus, if disorder wereto decrease in the contracting phase of the universe, one might also expect itto decrease inside a black hole. So perhaps an astronaut who fell into a blackhole would be able to make money at roulette by remembering where the ballwent before he placed his bet. Unfortunately, however, he would not have longto play before he was turned to spaghetti by the very strong gravitational fields.Nor would he be able to let us know about the reversal of the thermodynamicarrow, or even bank his winnings, because he would be trapped behind theevent horizon of the black hole.

At first, I believed that disorder would decrease when the universe recollapsed.This was because I thought that the universe had to return to a smooth andordered state when it became small again. This would have meant that thecontracting phase was like the time reverse of the expanding phase. People inthe contracting phase would live their lives backward. They would die beforethey were born and would get younger as the universe contracted. This idea isattractive because it would mean a nice symmetry between the expanding andcontracting phases. However, one cannot adopt it on its own, independent ofother ideas about the universe. The question is: Is it implied by the no bound-ary condition or is it inconsistent with that condition?

As I mentioned, I thought at first that the no boundary condition did indeedimply that disorder would decrease in the contracting phase. This was basedon work on a simple model of

the universe in which the collapsing phaselooked like the time reverse of the expanding phase. However, a colleague ofmine, Don Page, pointed out that the no boundary condition did not requirethe contracting phase necessarily to be the time reverse of the expandingphase. Further, one of my students, Raymond Laflamme, found that in a slightlymore complicated model, the collapse of the universe was very different fromthe expansion. I realized that I had made a mistake. In fact, the no boundarycondition implied that disorder would continue to increase during the con-traction. The thermodynamic and psychological arrows of time would notreverse when the universe begins to recontract or inside black holes.What should you do when you find you have made a mistake like that? Somepeople, like Eddington, never admit that they are wrong. They continue tofind new, and often mutually inconsistent, arguments to support their case.Others claim to have never really supported the incorrect view in the firstplace or, if they did, it was only to show that it was inconsistent. I could givea large number of examples of this, but I won't because it would make me toounpopular. It seems to me much better and less confusing if you admit in printthat you were wrong. A good example of this was Einstein, who said that thecosmological constant, which he introduced when he was trying to make astatic model of the universe, was the biggest mistake of his life.

The Theory of Everything: The Origin and Fate of the Universe

Chapter 7 - SEVENTH LECTURE - THE THEORY OF EVERYTHING

It would be very difficult to construct a complete unified theory of everythingall at one go. So instead we have made progress by finding partial theories.These describe a limited range of happenings and neglect other effects, orapproximate them by certain numbers. In chemistry, for example, we can cal-culate the interactions of atoms without knowing the internal structure of thenucleus of an atom. Ultimately, however, one would hope to find a complete,consistent, unified theory that would include all these partial theories asapproximations. The quest for such a theory is known as “the unification ofphysics.”

Einstein spent most of his later years unsuccessfully searching for a unified the-ory, but the time was not ripe: Very little was known about the nuclear forces.Moreover, Einstein refused to believe in the reality of quantum mechanics,despite the important role he had played in its development. Yet it seems thatthe uncertainty principle is a fundamental feature of the universe we live in. Asuccessful unified theory must therefore necessarily incorporate this principle.The prospects for finding such a theory seem to be much better now becausewe know so much more about the universe. But we must beware of overconfi-dence. We have had false dawns before. At the beginning of this century, forexample, it was thought that everything could be explained in terms of theproperties of continuous matter, such as elasticity and heat conduction. Thediscovery of atomic structure and the uncertainty principle put an end to that.Then again, in 1928, Max Born told a group of visitors to GöttingenUniversity, “Physics, as we know it, will be over in six months.” His confidencewas based on the recent discovery by Dirac of the equation that governed theelectron. It was thought that a similar equation would govern the proton,which was the only other particle known at the time, and that would be theend of theoretical physics. However, the discovery of the neutron and ofnuclear forces knocked that one on the head, too.

Having said this, I still believe there are grounds for cautious optimism that wemay now be near the end of the search for the ultimate laws of nature. At themoment, we have a number of partial theories. We have general relativity, thepartial theory of gravity, and the partial theories that govern the weak, thestrong, and the electromagnetic forces. The last three may be combined

inso-called grand unified theories. These are not very satisfactory because theydo not include gravity. The main difficulty in finding a theory that unifiesgravity with the other forces is that general relativity is a classical theory. Thatis, it does not incorporate the uncertainty principle of quantum mechanics. Onthe other hand, the other partial theories depend on quantum mechanics in anessential way. A necessary first step, therefore, is to combine general relativitywith the uncertainty principle. As we have seen, this can produce someremarkable consequences, such as black holes not being black, and the uni-verse being completely self-contained and without boundary. The trouble is,the uncertainty principle means that even empty space is filled with pairs ofvirtual particles and antiparticles. These pairs would have an infinite amountof energy. This means that their gravitational attraction would curve up theuniverse to an infinitely small size.

Rather similar, seemingly absurd infinities occur in the other quantum theories.However, in these other theories, the infinities can be canceled out by a processcalled renormalization. This involves adjusting the masses of the particles andthe strengths of the forces in the theory by an infinite amount. Although thistechnique is rather dubious mathematically, it does seem to work in practice. Ithas been used to make predictions that agree with observations to an extraor-dinary degree of accuracy. Renormalization, however, has a serious drawbackfrom the point of view of trying to find a complete theory. When you subtractinfinity from infinity, the answer can be anything you want. This means thatthe actual values of the masses and the strengths of the forces cannot bepredicted from the theory. Instead, they have to be chosen to fit the observa-tions. In the case of general relativity, there are only two quantities that can beadjusted: the strength of gravity and the value of the cosmological constant. Butadjusting these is not sufficient to remove all the infinities. One therefore hasa theory that seems to predict that certain quantities, such as the curvature ofspace-time, are really infinite, yet these quantities can be observed andmeasured to be perfectly finite. In an attempt to overcome this problem, a the-ory called “supergravity” was suggested in 1976. This theory was really just gen-eral relativity with some additional particles.

In general relativity, the gravitational force can be thought of as being carriedby a particle of spin 2 called the graviton. The idea was to add certain othernew particles of spin 3/2, 1, 1/2, and 0. In a sense, all these particles could thenbe regarded as different aspects of the same “superparticle.” The virtual parti-

cle/antiparticle pairs of spin 1/2 and 3/2 would have negative energy. Thiswould tend to cancel out the positive energy of the virtual pairs of particles ofspin 0, 1, and 2. In this way, many of the possible infinities would cancel out,but it was suspected that some infinities might still remain. However, the cal-culations required to find out whether there were any infinities left uncanceledwere so long and difficult that no one was prepared to undertake them. Evenwith a computer it was reckoned it would take at least four years. The chanceswere very high that one would make at least one mistake, and probably more.So one would know one had the right answer only if someone else repeated thecalculation and got the same answer, and that did not seem very likely.Because of this problem, there was a change of opinion in favor of what arecalled string theories. In these theories the basic objects are not particles thatoccupy a single point of space. Rather, they are things that have a length butno other dimension, like an infinitely thin loop of string. A particle occupiesone point of space at each instant of time. Thus, its history can be representedby a line in space-time called the "world-line." A string, on the other hand,occupies a line in space at each moment of time. So its history in space-timeis a two-dimensional surface called the "world-sheet." Any point on such aworld-sheet can be described by two numbers, one specifying the time and theother the position of the point on the string. The world-sheet of a string is acylinder or tube. A slice through the tube is a circle, which represents the posi-tion of the string at one particular time.

Two pieces of string can join together to form a single string. It is like the twolegs joining on a pair of trousers. Similarly, a single piece of string can divideinto two strings. In string theories, what were previously thought of as particlesare now pictured as waves traveling down the string, like waves on a washingline. The emission or absorption of one particle by another corresponds to thedividing or joining together of strings. For example, the gravitational force ofthe sun on the Earth corresponds to an H-shaped tube or pipe. String theory israther like plumbing, in a way. Waves on the two vertical sides of the H corre-spond to the particles in the sun and the Earth, and waves on the horizontalcrossbar correspond to the gravitational force that travels between them.String theory has a curious history. It was originally invented in the late 1960sin an attempt to find a theory to describe the strong force. The idea was thatparticles like the proton and the neutron could be regarded as waves on astring. The strong forces between the particles would correspond to

pieces ofstring that went between other bits of string, like in a spider's web. For this the-ory to give the observed value of the strong force between particles, the stringshad to be like rubber bands with a pull of about ten tons.

In 1974 Joël Scherk and John Schwarz published a paper in which they showedthat string theory could describe the gravitational force, but only if the tensionin the string were very much higher-about 1039tons. The predictions of thestring theory would be just the same as those of general relativity on normallength scales, but they would differ at very small distances-less than 10-33centimeters. Their work did not receive much attention, however, because atjust about that time, most people abandoned the original string theory of thestrong force. Scherk died in tragic circumstances. He suffered from diabetesand went into a coma when no one was around to give him an injection ofinsulin. So Schwarz was left alone as almost the only supporter of stringtheory, but now with a much higher proposed value of the string tension.There seemed to have been two reasons for the sudden revival of interest instrings in 1984. One was that people were not really making much progresstoward showing that supergravity was finite or that it could explain the kindsof particles that we observe. The other was the publication of a paper by JohnSchwarz and Mike Green which showed that string theory might be able toexplain the existence of particles that have a built-in left-handedness, likesome of the particles that we observe. Whatever the reasons, a large numberof people soon began to work on string theory. A new version was developed,the so-called heterotic string. This seemed as if it might be able to explain thetypes of particle that we observe.

String theories also lead to infinities, but it is thought they will all cancel outin versions like the heterotic string. String theories, however, have a biggerproblem. They seem to be consistent only if space-time has either ten ortwenty-six dimensions, instead of the usual four. Of course, extra space-timedimensions are a commonplace of science fiction; indeed, they are almost anecessity. Otherwise, the fact that relativity implies that one cannot travelfaster than light means that it would take far too long to get across our owngalaxy, let alone to travel to other galaxies. The science fiction idea is that onecan take a shortcut through a higher dimension. One can picture this in thefollowing way. Imagine that the space we live in had only two dimensions andwas curved like the surface of a doughnut or a torus. If you were on one side ofthe ring and you wanted to get to a point on the other side, you would have togo around the

ring. However, if you were able to travel in the third dimension,you could cut straight across.

Why don't we notice all these extra dimensions if they are really there? Whydo we see only three space and one time dimension? The suggestion is that theother dimensions are curved up into a space of very small size, something likea million million million million millionth of an inch. This is so small that wejust don't notice it. We see only the three space and one time dimension inwhich space-time is thoroughly flat. It is like the surface of an orange: if youlook at it close up, it is all curved and wrinkled, but if you look at it from adistance, you don't see the bumps and it appears to be smooth. So it is withspace-time. On a very small scale, it is ten-dimensional and highly curved.But on bigger scales, you don't see the curvature or the extra dimensions.If this picture is correct, it spells bad news for would-be space travelers. Theextra dimensions would be far too small to allow a spaceship through.However, it raises another major problem. Why should some, but not all, ofthe dimensions be curled up into a small ball? Presumably, in the very earlyuniverse, all the dimensions would have been very curved. Why did threespace and one time dimension flatten out, while the other dimensionsremained tightly curled up?

One possible answer is the anthropic principle. Two space dimensions do notseem to be enough to allow for the development of complicated beings like us.For example, two-dimensional people living on a one-dimensional Earthwould have to climb over each other in order to get past each other. If a two-dimensional creature ate something it could not digest completely, it wouldhave to bring up the remains the same way it swallowed them, because if therewere a passage through its body, it would divide the creature into two separateparts. Our two-dimensional being would fall apart. Similarly, it is difficult tosee how there could be any circulation of the blood in a two-dimensional crea-ture. There would also be problems with more than three space dimensions.The gravitational force between two bodies would decrease more rapidly withdistance than it does in three dimensions. The significance of this is that theorbits of planets, like the Earth, around the sun would be unstable. The leastdisturbance from a circular orbit, such as would be caused by the gravitationalattraction of other planets, would cause the Earth to spiral away from or intothe sun. We would either freeze or be burned up. In fact, the same behavior ofgravity with distance would mean that the sun would also be unstable. It wouldeither fall apart or it would collapse to form a black hole. In either case,

itwould not be much use as a source of heat and light for life on Earth. On asmaller scale, the electrical forces that cause the electrons to orbit around thenucleus in an atom would behave in the same way as the gravitational forces.Thus, the electrons would either escape from the atom altogether or it would spi-ral into the nucleus. In either case, one could not have atoms as we know them.It seems clear that life, at least as we know it, can exist only in regions ofspace-time in which three space and one time dimension are not curled upsmall. This would mean that one could appeal to the anthropic principle, pro-vided one could show that string theory does at least allow there to be suchregions of the universe. And it seems that indeed each string theory doesallow such regions. There may well be other regions of the universe, or otheruniverses (whatever that may mean) in which all the dimensions are curledup small, or in which more than four dimensions are nearly flat. But therewould be no intelligent beings in such regions to observe the different num-ber of effective dimensions.

Apart from the question of the number of dimensions that space-time appearsto have, string theory still has several other problems that must be solvedbefore it can be acclaimed as the ultimate unified theory of physics. We do notyet know whether all the infinities cancel each other out, or exactly how torelate the waves on the string to the particular types of particle that weobserve. Nevertheless, it is likely that answers to these questions will be foundover the next few years, and that by the end of the century we shall knowwhether string theory is indeed the long sought-after unified theory of physics.Can there really be a unified theory of everything? Or are we just chasing amirage? There seem to be three possibilities:

There really is a complete unified theory, which we will somedaydiscover if we are smart enough.

There is no ultimate theory of the universe, just an infinitesequence of theories that describe the universe more and moreaccurately.

There is no theory of the universe. Events cannot be predictedbeyond a certain extent but occur in a random and arbitrary manner.Some would argue for the third possibility on the grounds that if there were acomplete set of laws, that would infringe on God's freedom to change His mindand to intervene in the world. It's a bit like the old paradox: Can God make astone so heavy that He can't lift it? But the idea that God might want tochange His mind is an example of the fallacy, pointed out by St. Augustine, ofimagining God as a being existing in time. Time is a

property only of theuniverse that God created. Presumably, He knew what He intended when Heset it up.

With the advent of quantum mechanics, we have come to realize that eventscannot be predicted with complete accuracy but that there is always a degreeof uncertainty. If one liked, one could ascribe this randomness to the interven-tion of God. But it would be a very strange kind of intervention. There is noevidence that it is directed toward any purpose. Indeed, if it were, it wouldn'tbe random. In modern times, we have effectively removed the third possibilityby redefining the goal of science. Our aim is to formulate a set of laws that willenable us to predict events up to the limit set by the uncertainty principle.The second possibility, that there is an infinite sequence of more and morerefined theories, is in agreement with all our experience so far. On many occa-sions, we have increased the sensitivity of our measurements or made a newclass of observations only to discover new phenomena that were not predictedby the existing theory. To account for these, we have had to develop a moreadvanced theory. It would therefore not be very surprising if we find that ourpresent grand unified theories break down when we test them on bigger andmore powerful particle accelerators. Indeed, if we didn't expect them to breakdown, there wouldn't be much point in spending all that money on buildingmore powerful machines.

However, it seems that gravity may provide a limit to this sequence of "boxeswithin boxes." If one had a particle with an energy above what is called thePlanck energy, 1019 GeV, its mass would be so concentrated that it would cutitself off from the rest of the universe and form a little black hole. Thus, it doesseem that the sequence of more and more refined theories should have somelimit as we go to higher and higher energies. There should be some ultimatetheory of the universe. Of course, the Planck energy is a very long way fromthe energies of around a GeV, which are the most that we can produce in thelaboratory at the present time. To bridge that gap would require a particleaccelerator that was bigger than the solar system. Such an accelerator wouldbe unlikely to be funded in the present economic climate.However, the very early stages of the universe are an arena where such ener-gies must have occurred. I think that there is a good chance that the study ofthe early universe and the requirements of mathematical consistency will leadus to a complete unified theory by the end of the century-always presumingwe don't blow ourselves up first.

What would it mean if we actually did discover the ultimate

theory of the uni-verse? It would bring to an end a long and glorious chapter in the history ofour struggle to understand the universe. But it would also revolutionize theordinary person's understanding of the laws that govern the universe. InNewton's time it was possible for an educated person to have a grasp of thewhole of human knowledge, at least in outline. But ever since then, the paceof development of science has made this impossible. Theories were alwaysbeing changed to account for new observations. They were never properlydigested or simplified so that ordinary people could understand them.You hadto be a specialist, and even then you could only hope to have a proper grasp ofa small proportion of the scientific theories.Further, the rate of progress was so rapid that what one learned at school oruniversity was always a bit out of date. Only a few people could keep upwith the rapidly advancing frontier of knowledge. And they had to devotetheir whole time to it and specialize in a small area. The rest of the popula-tion had little idea of the advances that were being made or the excitementthey were generating.

Seventy years ago, if Eddington is to be believed, only two people understoodthe general theory of relativity. Nowadays tens of thousands of university grad-uates understand it, and many millions of people are at least familiar with theidea. If a complete unified theory were discovered, it would be only a matterof time before it was digested and simplified in the same way. It could then betaught in schools, at least in outline. We would then all be able to have someunderstanding of the laws that govern the universe and which are responsiblefor our existence.

Einstein once asked a question: "How much choice did God have in construct-ing the universe?" If the no boundary proposal is correct, He had no freedomat all to choose initial conditions. He would, of course, still have had the free-dom to choose the laws that the universe obeyed. This, however, may notreally have been all that much of a choice. There may well be only one or asmall number of complete unified theories that are self-consistent and whichallow the existence of intelligent beings.

We can ask about the nature of God even if there is only one possible unifiedtheory that is just a set of rules and equations. What is it that breathes fire intothe equations and makes a universe for them to describe? The usual approachof science of constructing a mathematical model cannot answer the questionof why there should be a universe for the model to describe. Why does the uni-verse go to all the bother of existing? Is the unified theory so compelling thatit brings about its own existence? Or does it need a creator, and, if so, does Hehave any effect on the

universe other than being responsible for its existence?And who created Him?

Up until now, most scientists have been too occupied with the developmentof new theories that describe what the universe is, to ask the question why. Onthe other hand, the people whose business it is to ask why-the philoso-phers-have not been able to keep up with the advance of scientific theories.In the eighteenth century, philosophers considered the whole of humanknowledge, including science, to be their field. They discussed questions suchas: Did the universe have a beginning? However, in the nineteenth and twen-tieth centuries, science became too technical and mathematical for thephilosophers or anyone else, except a few specialists. Philosophers reduced thescope of their inquiries so much that Wittgenstein, the most famous philoso-pher of this century, said, “The sole remaining task for philosophy is the analy-sis of language.” What a comedown from the great tradition of philosophyfrom Aristotle to Kant.

However, if we do discover a complete theory, it should in time be understand-able in broad principle by everyone, not just a few scientists. Then we shall allbe able to take part in the discussion of why the universe exists. If we find theanswer to that, it would be the ultimate triumph of human reason. For then wewould know the mind of God.